МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В. ЛОМОНОСОВА

ФАКУЛЬТЕТ ФУНДАМЕНТАЛЬНОЙ МЕДИЦИНЫ

На правах рукописи

Арбатский Михаил Спартакович

Выяснение механизмов развития гетерогенного ответа мезенхимных стромальных клеток на профибротические стимулы с использованием анализа транскриптома единичных клеток

Специальность 1.5.8 Математическая биология, биоинформатика

ДИССЕРТАЦИЯ

на соискание ученой степени

кандидата биологических наук

Научный руководитель:

доктор медицинских наук, доцент

Ефименко Анастасия Юрьевна

Москва 2026

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность и степень разработанности темы исследования

С момента появления принципиально новых возможностей в анализе транскриптомных данных биологических образцов на уровне единичных клеток начало развиваться новое направление в биоинформатике – анализ данных РНК секвенирования единичных клеток (scRNA-seq). На текущий момент сформирована необходимая инфраструктура для этого направления, включая наборы реактивов, платформы для подготовки библиотек, библиотеки пакетов и программные конвейеры для процессинга данных. Особый интерес в изучении нового направления в биоинформатике представляет постоянное развитие методов обработки данных scRNA-seq. После освоения основных возможностей новой технологии продолжают появляться принципиально новые подходы к анализу данных scRNA-seq. Это связано с существенно более сложной структурой и объемом данных.

Разработка новой технологии была продиктована актуальной проблемой определения типа отдельных клеток или их субпопуляций, а также подробного описания молекулярных процессов в отдельных клетках. Внедрение scRNA-seq позволило получить ценную информацию о новых субпопуляциях клеток, присутствие которых в тканях определяет патогенез таких актуальных для здравоохранения состояний, как онкологические заболевания, метаболические нарушения и фибротические изменения органов и тканей.

Способность тканей млекопитающих, в том числе человека, восстанавливать свою структуру и функции после повреждения обеспечивается за счет репаративной регенерации, однако при длящемся или обширном повреждении, сопровождающемся развитием хронического воспаления, исходом этого процесса часто является возникновение фиброза. При этом происходит избыточное отложение белков внеклеточного матрикса (ВКМ) и его ремоделирование, что приводит к замещению функциональной

ткани соединительной и в конечном итоге к дисфункции органов и тканей. Считается, что основную роль в этих процессах играют миофибробласты, которые дифференцируются под действием профибротических сигналов из фибробластов и других типов клеток стромы.

Анализ имеющихся к настоящему времени данных свидетельствует о том, что мезенхимные стромальные клетки (МСК) принимают участие в регуляции практически всех физиологических процессов, происходящих при репарации и регенерации ткани, в том числе вовлечены в регуляцию фиброза. При этом основная роль в реализации их регуляторных эффектов отводится биологически активным компонентам секретома МСК, включая пул некодирующих регуляторных РНК (нкРНК), которые будучи перенесенными в составе внеклеточных везикул в другие клетки, способны перепрограммировать их в направлении стимуляции регенеративных процессов. Это делает МСК одним из наиболее перспективных инструментов регенеративной медицины. Однако показано, что в ответ на различные сигналы от поврежденных тканей МСК могут сами дифференцироваться в миофибробласты, а также реагировать специфическими изменениями своего секретома (Grigorieva et al., 2024). При этом остается открытым вопрос о наличии субпопуляций МСК, по-разному отвечающих на профибротические сигналы, и участия в этих процессах нкРНК, секретируемых МСК

Цели и задачи исследования

Целью работы является установление механизмов гетерогенности ответа МСК на сигналы профибротического микроокружения с помощью метода scRNA-seq.

Для достижения цели были сформулированы следующие задачи:

1) Выявить субпопуляции МСК, различающиеся по транскриптомным паттернам при культивировании в модели профибротического микроокружения, с помощью метода scRNA-seq.

2) Проанализировать особенности транскриптомного паттерна субпопуляций МСК, которые дифференцируются и не дифференцируются в миофибробласты, идентифицировать в них клеточные типы с использованием ручных и автоматизированных методов типирования и провести анализ регулонов в выявленных кластерах с помощью метода получения регуляторных сетей с целью предположить функциональные характеристики различающихся по транскриптомным паттернам субпопуляций.

3) Построить траектории для определения направления развития МСК, культивируемых в профибротических условиях, с помощью метода упорядочивания единичных клеток вдоль псевдовремени.

4) Установить особенности экспрессии нкРНК в субпопуляциях МСК, по-разному отвечающих на профибротические сигналы.

5) Выявить мембранный белок, экспрессия которого значимо различается в субпопуляциях МСК, по-разному отвечающих на профибротические сигналы, с целью выбора маркера для разделения субпопуляций с использованием клеточного сортинга.

Объект и предмет исследования

Объектом исследования являются мезенхимные стромальные клетки (МСК) человека, опосредующие поддержание клеточного состава тканей в течение всей жизни организма, в том числе и после повреждений. Такие клетки, их функционально специализированные субпопуляции, а также продуцируемые МСК биологически активные вещества, включая белки, липиды и нуклеиновые кислоты, могут быть использованы для разработки новых подходов к терапии различных заболеваний.

Предметом исследования являются транскриптомные данные scRNA-seq, полученные в результате секвенирования подготовленных библиотек МСК.

Научная новизна

В работе впервые проведен анализ результатов scRNA-seq МСК человека, культивированных в стандартных условиях и в условиях профибротического микроокружения, с использованием современных биоинформатических методов. Предложены новые подходы к идентификации типов клеток в кластерах, находящихся на линии траекторий развития, исходя из знаний о клеточных дифферонах. Впервые было установлено, что в популяции МСК присутствуют субпопуляции, по-разному отвечающие на профибротические сигналы. Детальная характеристика субпопуляции МСК, которая не дифференцируется в этих условиях в миофибробласты, по дифференциально-экспрессирующимся генам, биологическим процессам и регулонам, позволила сформировать гипотезы о функциональных свойствах этих клеток. Для выявленной субпопуляции установлены характерные поверхностные маркеры, что позволило отсортировать эти клетки для дальнейшего экспериментального изучения и валидации выявленных отличий в их свойствах. Впервые, с помощью метода построения траекторий развития клеточных популяций, для МСК описаны возможные направления их дифференцировки под действием профибротических стимулов, что может быть использовано для контроля дифференцировки МСК в заданном направлении. Весомым дополнением работы стал анализ предшественников микроРНК в выявленных субпопуляциях, что позволило спрогнозировать их участие в регуляции фиброза.

Теоретическая и практическая значимость

Применение новых методов биоинформатической обработки результатов анализа scRNA-seq дает уникальные преимущества для изучения гетерогенности популяции клеток, что крайне важно для понимания особенностей ответа клеток на различные стимулы. Результаты диссертационного исследования позволили установить разнообразие ответов

МСК в модели, воспроизводящей профибротическое микроокружение. При кластеризации выявлена субпопуляция МСК с потенциально антифибротическими свойствами и проведена ее характеристика. Практической значимостью работы является получение данных, обосновывающих необходимость изучения этой субпопуляции с целью выявления механизмов участия этих клеток в развитии фиброза. Контроль над процессом дифференцировки МСК в фибробласты и миофибробласты, стимуляция регуляторных свойств изучаемой субпопуляции и ее влияния на миофибробласты и фибробласты может позволить предотвращать и обращать развитие фибротических процессов в организме. Новизна и практическая значимость работы подтверждаются полученным патентом РФ №2766707 от 15.03.2022 г. «Средство для лечения фиброза тканей на основе компонентов секретома мезенхимных стромальных клеток, способ получения и применения средства».

Возможное терапевтическое использование

Быстрый, нацеленный биоинформатический анализ данных секвенирования биологического образца ткани, взятой у пациента, может помочь спрогнозировать поведение соединительной ткани и реакции на различные виды стимулов. В частности механические. Такой подход может быть актуален в эстетической челюстно-лицевой хирургии, где выраженность фибротических процессов может привести к нежелательным для пациента последствиям, связанным с деформацией форм оперируемых областей.

В случае выявления агрессивного характера реакции соединительной ткани возможно применение клеточных продуктов МСК, как фактора, регулирующего патологический процесс ранозаживления с преимущественным образованием соединительной ткани без восстановления ее функции. В частности, использование внеклеточных везикул МСК, в качестве БМКП может повлиять на дифференцировку фибробластов в

миофибробласты, снижая выраженность этого направления. Также, при выявлении достаточно представленной субпопуляции клеток, сдерживающих развитие фибротических изменений, возможна стимуляция этих клеток с целью снижения скорости непродуктивной репарации ткани.

Методология и методы исследования

Работа была выполнена с использованием разнообразных программ и пакетов, а также программных сценариев, написанных самостоятельно.

Для выравнивания, квантификации, уменьшения размерности, кластеризации, расчета дифференциальной экспрессии, получения файлов barcodes.tsv, features.tsv и matrix.mtx и .cloupe-файлов из первичных данных в формате .fastq был использован cellranger count (v.7). Для получения .aggr-файлов использовался cellranger aggr. Для визуализации и анализа .cloupe-файлов использовалась программа Loupe Browser.

Для создания Seurat-объекта из barcodes.tsv, features.tsv и matrix.mtx был использован R-пакет Seurat. Сравнивались результаты таких пайплайнов для анализа данных scRNA-seq, как SCANPY и Asc-Seurat.

Интеграция образцов в режимах CCA, RPCA и SCTransform производилась с помощью R-пакета Seurat.

Автоматическое типирование клеток образцов проводилось с помощью R-пакетов SingleR и celldex. Ручное типирование осуществлялось с использованием клеточных референсов PanglaoDB и CellMarker.

Получение .loom-файлов реализовано с помощью пакета velocyto, существующего в двух форматах Python и R. Получение траекторий развития клеток образцов осуществлялось с помощью Python-пакета scVelo, коллекции R-пакетов Dynverse и веб-платформы Asc-Seurat.

Для получения графиков RNA-velocity использовался Python-пакет scVelo.

Для работы с данными использовались среды разработки RStudio, Google Colab и PyCharm.

Положения, выносимые на защиту

1) Факторы профибротического микроокружения могут определять гетерогенность популяции МСК, реализующейся в формировании клеточных субпопуляций с различными транкриптомными паттернами.

2) С помощью анализа данных РНК секвенирования единичных клеток установлено, что в модели профибротического микроокружения только часть МСК дифференцируется в миофибробласты, остальные клетки обладают транскриптомным профилем, отличным от характерного для миофибробластов.

3) Субпопуляция МСК, не дифференцирующаяся в миофибробласты под действием профибротических стимулов, характеризуется повышенной экспрессией групп генов, отвечающих за организацию и ремоделирование ВКМ, регуляцию метаболических процессов и ангиогенеза, и может быть отделена от других субпопуляций по экспрессии PDGFRa.

4) Гетерогенность ответов МСК на профибротические стимулы может оказывать влияние на развитие фиброза тканей за счет регуляции баланса между субпопуляциями клеток, пополняющих пул миофибробластов, и клеток с предположительно антифибротическими свойствами.

Личный вклад автора

Основные результаты, представленные в диссертационной работе, получены лично автором. Его вклад включает анализ научной литературы, разработку новых вычислительных методов, планирование и проведение вычислительных экспериментов, обработку и интерпретацию данных, подготовку публикаций и участие в научных конференциях. В работе O. Grigorieva et al. 2024 автором была валидирована in vitro модель профибротического микроокружения на основе децеллюляризованного

внеклеточного матрикса и TGFβ-1. В исследовании N. Basalova et al. 2023 им был проведен биоинформатический анализ данных секвенирования внеклеточных везикул МСК и выявлены ключевые микроРНК (miR-29c и miR-129), участвующие в регуляции фиброза. В работе A.A. Khozyainova et al. 2023 автор систематизировал методы биоинформатической обработки данных scRNA-seq, включая контроль качества, фильтрацию, кластеризацию и анализ траекторий развития клеточных популяций. В работе O. Grigorieva et al. 2023 автором был выполнен анализ данных scRNA-seq для выявления регуляторных генов дифференцировки МСК в миофибробласты, что позволило установить роль CHD3 и RDH10 в этом процессе. В исследовании M. Arbatsky et al. 2022 им был разработан подход к многомерной визуализации данных методом главных компонент для более полного анализа транскриптомных данных. В работах N. Alexandrushkina et al. 2020 и N. Basalova et al. 2020 автор провел биоинформатический анализ данных bulkRNA-seq и секвенирования внеклеточных везикул для выявления молекулярных механизмов антифибротического действия МСК. В патенте № 2766707 C1 «Средство для лечения фиброза тканей на основе компонентов секретома мезенхимных стромальных клеток, способ получения и применения средства» личный вклад Арбатского М.С. заключался в применении биоинформатических подходов для анализа компонентов секретома мезенхимальных стромальных клеток и прогнозирования их терапевтического потенциала при фиброзе тканей.

Степень достоверности данных

Данные, представленные в работе, получены с использованием современных программ и пакетов. Результаты воспроизводимы. Обзор литературы и обсуждение подготовлены с использованием актуальной литературы.

Публикации по теме диссертации

По теме диссертации опубликовано 6 печатных работ в рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности 1.5.8 Математическая биология, биоинформатика.

1. Grigorieva O., Basalova N., Dyachkova U., Novoseletskaya E., Vigovskii M., **Arbatskiy M.**, Kulebyakina M., Efimenko A. Modeling the profibrotic microenvironment in vitro: model validation // **Biochemical and Biophysical Research Communications. – 2024**. Vol. 733. P. 150574. EDN: HQSAIE, Импакт-фактор 2,2 (JIF). (0,84/0,15)[1].

2. Basalova N., **Arbatskiy M.**, Popov V., Grigorieva O., Vigovskiy M., Zaytsev I., Novoseletskaya E., Sagaradze G., Danilova N., Malkov P., Cherniaev A., Samsonova M., Karagyaur M., Tolstoluzhinskaya A., Dyachkova U., Akopyan Z., Tkachuk V., Kalinina N., Efimenko A. Mesenchymal stromal cells facilitate resolution of pulmonary fibrosis by miR-29c and miR-129 intercellular transfer // **Experimental and Molecular Medicine. – 2023**. – Vol. 55. – № 7. – pp. 1399–1412. EDN: KSCMMI. Импакт-фактор 12,9 (JIF). (0,78/0,1).

3. Khozyainova A.A., Valyaeva A.A., Arbatsky M.S., Isaev S.V., Iamshchikov P.S., Volchkov E.V., Sabirov M.S., Zainullina V.R., Chechekhin V.I., Vorobev R.S., Menyailo M.E., Tyurin-Kuzmin P.A., Denisov E.V. Complex Analysis of Single-Cell RNA Sequencing Data // **Biochemistry (Moscow)** – **2023. –** Vol. 88. – № 2. – pp. 231–252. EDN: QFSJMW. Импакт-фактор 2,2 (JIF). (1,26/0,3).

4. Grigorieva O., Basalova N., Vigovskiy M., **Arbatskiy M.,** Dyachkova U., Kulebyakina M., Kulebyakin K., Tyurin-Kuzmin P., Kalinina N., Efimenko A. Novel Potential Markers of Myofibroblast Differentiation Revealed by Single-Cell RNA Sequencing Analysis of Mesenchymal Stromal

[1] В скобках приведён объём публикации в печатных листах и вклад автора в печатных листах

Cells in Profibrotic and Adipogenic Conditions // **Biomedicines.** – **2023**. – Vol. 11. – № 3. – P. 840. doi: 10.3390/biomedicines11030840. EDN: OUZBFK. Импакт-фактор 3,9 (JIF). (0,96/0,15).

5. **Arbatsky M.**, Tyurin-Kuzmin P., Kulebyakin K., Chechekhin V., Kalinina N., Sysoeva V., Semina E., Rubina K. Points of Significance: Principal Component Analysis for Biocentric Data Visualization // **BioNanoScience. – 2022. –** Vol. 12. – pp. 1366–1380. doi: 10.1007/s12668-022-01021-w. EDN: QAKCPE. Импакт-фактор 3,2 (JIF). (0,84/0,2).

6. Basalova N., Sagaradze G., **Arbatskiy M.,** Evtushenko E., Kulebyakin K., Grigorieva O., Akopyan Z., Kalinina N., Efimenko A. Secretome of Mesenchymal Stromal Cells Prevents Myofibroblasts Differentiation by Transferring Fibrosis-Associated microRNAs within Extracellular Vesicles // **Cells. – 2020**. – Vol. 9. – №5. – P. 1272, EDN: ESMODC. Импакт-фактор 5,2 (JIF). (0,84/0,2).

Теоретические и практические результаты использовались в выполнении задач в качестве исполнителя по грантам РФФИ «Роль нестин экспрессирующих мезенхимных мультипотентных клеток в адипогенезе», «Участие некодирующих регуляторных РНК, секретируемых мезенхимными стромальными клетками, в процессах регенерации и репарации тканей» и гранту РНФ «Фундаментальные проблемы регенеративной медицины: регуляция обновления и репарации тканей человека».

Апробация результатов

Основные результаты, положения и выводы диссертационного исследования были представлены на 19 научных конференциях и симпозиумах: конгрессе «Single Cells: Technology to Biology» (Сингапур, 24-26 февраля 2019 г.), International Congress Biotechnology: State of the art and perspectives (Москва, 25-27 февраля 2019 г.), Regulatory RNAs (Берлин, 12-14 мая 2019 г.), TERMIS-EU 2019 (Родос, 27-31 мая 2019 г.), IV Национальный

конгресс по регенеративной медицине (Москва, 20-23 ноября 2019 г.), VII Молодёжная школа-конференция по молекулярной и клеточной биологии Института цитологии РАН (Санкт-Петербург, 12-15 октября 2020 г.), ISEV 2021 (Франция, 18-21 мая 2021 г.), The 45th FEBS Congress (Словения, 3-8 июля 2021 г.), Noncoding RNA World: From Mechanism to Therapy (Швейцария, 21-23 июля 2021 г.), Moscow Conference on Computational Molecular Biology (MCCMB) (Москва, 30 июля – 2 августа 2021 г.), III Научно-практическая конференция "Секвенирование единичных клеток" (Томск, 23-27 августа 2021 г.), ESGCT Collaborative Virtual Congress 2021 (Бельгия, 19-22 октября 2021 г.), 25-ая Пущинская школа-конференция молодых ученых с международным участием «Биология – наука XXI века» (Пущино, 18-22 апреля, 2022 г.), IV Научно-практическая школа «Анализ отдельных клеток» (Томск, Томский национальный исследовательский медицинский центр Российской академии наук, 22-26 августа 2022 г.), III Объединенный Научный Форум Физиологов, Биохимиков И Молекулярных Биологов (Сочи, Россия, 3-7 октября 2022 г.), VIII Молодёжная Школа-Конференция по молекулярной биологии и генетическим технологиям Института цитологии РАН (Санкт-Петербург, Россия, 10-14 октября 2022 г.), V Национальный Конгресс по Регенеративной Медицине (Москва, МГУ имени М.В. Ломоносова, 23–25 ноября 2022 г.), 3-я международная конференция "Системная биология и системная физиология: регуляция сложных биологических систем" Внутриклеточная сигнализация и регуляция метаболизма (Москва, Россия, 2-4 декабря 2022), II Международный онлайн-конгресс «Управление старением» (Москва, Россия, 15-16 декабря 2022 г.), European Society of Human Genetics conference (ESHG 2023) (Глазго, United Kingdom, 10-13 июня 2023), Всемирный конгресс: Теория систем, алгебраическая биология, искусственный интеллект: математические основы и приложения (Москва, Россия, 26 июня - 30 августа 2023), 11-я Московская конференция по вычислительной молекулярной биологии (MCCMB) (Москва, Россия, 3-6 августа 2023), MCCMB 2023, Москва, Территория

Инновационного Центра “Сколково”, (Россия, 3-6 августа 2023), MCCMB 2023, Москва, Территория Инновационного Центра “Сколково” (Россия, 3-6 августа 2023), Вычислительная биология и искусственный интеллект для персонализированной медицины (Россия, 9-11 августа 2023), Конгресс “CRISPR-2023” (Новосибирск, Россия, 11-13 сентября 2023), TERMIS-AP 2023 (Гонконг, China, 16-19 октября 2023), IV Всероссийская научно-практическая конференция с международным участием «Развитие физико-химической биологии и биотехнологии на современном этапе» (Иркутск, Россия, 25-27 октября 2023), ESHG 2024 - European Human Genetics Conference (Берлин, Germany, 1-4 июня 2024), ESHG 2024 - European Human Genetics Conference (Берлин, Germany, 1-4 июня 2024), Вычислительная биология и искусственный интеллект для персонализированной медицины (Россия, 7-9 августа 2024).

На основании данных диссертации подготовлен и получен патент РФ № 2766707 от 15.03.2022 г.

Материалы диссертационного исследования также апробированы автором в преподавании учебных курсов «Биоинформатика и компьютерные технологии» и «Омиксные технологии» в рамках магистерской программы «Регенеративная биомедицина» по направлению подготовки 06.04.01 «Биология».

Соответствие диссертации паспорту научной специальности

Исследование проведено в рамках направлений исследований паспорта специальности 1.5.8 Математическая биология, биоинформатика: 2. Компьютерная системная биология (геномика, транскриптомика, протеомика, метаболомика, другие омиксные исследования) и 5. Идентификация потенциальных биомаркеров с целью диагностики заболеваний и перспективных молекулярных мишеней новых лекарств.

Структура и объем диссертации

Диссертация состоит из введения, трех глав, заключения, списка литературы, включающего 255 наименования, и двух приложений. Диссертация изложена на 223 страницах машинописного текста, содержит 49 рисунков, 26 таблиц.

Глава 1. ОБЗОР ЛИТЕРАТУРЫ

1.1 МСК. Определение, типирование, гетерогенность, участие в фиброзе

Изначально МСК (мезенхимные стромальные клетки) были обнаружены в кроветворных органах в виду колониеобразующих единиц фибробластов. МСК были охарактеризованы Фриденштейном как мультипотентные предшественники, которые могут дифференцироваться в адипоциты, остеобласты и хондробласты [1].

Под МСК в данном случае понимаются клетки, произошедшие из эмбриональной мезенхимной ткани, находящиеся в строме ткани или периваскулярной нише, которые вносят вклад в образование ВКМ, клеток соединительной ткани при хронических заболеваниях и острых травмах, а также поддерживающих тканевой гомеостаз. К таким клеткам относятся первичные периваскулярные стволовые клетки, комитированные предшественники клеток соединительной ткани и фибробласты.

Международным Обществом Генной и Клеточной Терапии (International Society for Cell and Gene Therapy, ISCT) был выработан ряд критериев, согласно которым можно охарактеризовать клетку как МСК:

1) Способность к быстрой адгезии;
2) Экспрессия поверхностных маркеров CD105, CD73, и CD90;
3) Отсутствие экспрессии маркеров CD45, CD34, CD14, CD11b, CD79α, CD19, и HLA-DR;
4) Способность дифференцироваться in vitro в остеобласты, хондробласты и адипоциты [2, 3].

Несмотря на то, что самым ранним источником этих клеток были клетки костного мозга, резидентные МСК были выявлены почти у всех эмбриональных и постнатальных тканях человека [4] МСК характеризуются способностью к самообновлению и разнонаправленной дифференцировкой in vitro, а также секрецией ростовых факторов и иммуномодулирующими

свойствами. Несмотря на отсутствие значительной теломеразной активности МСК можно выращивать в течение многих поколений in vitro, хотя долгое культивирование может вызвать старение МСК, привести к потере дифференцировочного потенциала и в конечном итоге к трансформации. В присутствии ростовых стимулирующих дифференцировку МСК они могут дифференцироваться в различных направлениях, включая адипоциты, остеобласты и хондробласты. Также было показано, что МСК модулируют созревание, пролиферацию, активацию, секрецию цитокинов и костимулирующие молекулы, экспрессирующиеся в различных иммунных клетках. По данным международного общества клеточной терапии минимальными требованиями для идентификации МСК являются – адгезия к пластику, дифференцировка в трех направлениях (адипоциты, остеобласты и хондробласты) и экспрессия набора поверхностных маркеров.

Из вышеприведенной информации можно сделать вывод, что МСК обладают идентичностью, связанной с тканью, в которой они находятся [5].

Доказательства, подтверждающие уникальную анатомическую идентичность фиброгенных стромальных клеток получены из недавнего исследования фенотипически сходных периваскулярных популяций МСК. Эти клетки имеют одинаковы набор маркеров (CD146+ CD45− CD34−) и, как ранее было обнаружено, обладают способностью давать начало нескольким соединительным тканям, и они представляют собой эквипотентную популяцию МСК [6]. Недавний анализ транскрипционных профилей и отслеживания перемещения клеток соединительной ткани показал, что (CD146+ CD45− CD34−) клетки находящиеся в костях и скелетных мышцах могут образовывать остеобласты и хондробласты in vivo, клетки находящиеся не в скелетных тканях не дифференцируются в этом направлении. Также, глубокий анализ транскриптома выявил существенные различия в экспрессии мембранных рецепторов и сигнальных молекул signal transducer and activator of transcription 2 (STAT2), transforming growth factor–β

receptor 2 (TGFβ2), fibroblast growth factor-18 (FGF-18), и retinoic acid receptor-α (RARA) были найдены в значительно большем колчестве в МСК из костного мозга по сравнению с МСК из других тканей, где экспрессия insulin-like growth factor 2 (IGF-2), jagged 1 (JAG-1), bone morphogenic protein 2 (BMP-2), FGF-13, and angiopoietin-like 1 (ANGPTL-1) была повышена в МСК из гладкомышечных волокон и TGFBR3, platelet-derived growth factor receptor-α (PDGFR-α), WNT1-inducible signaling protein 1 (WISP-1), interleukin-7 (IL-7), osteoglycin (OGN), IGF-1, and suppressor of cytokine signaling 5 (SOCS-5) были определены в МСК надкостницы [7].

Изучение МСК in vivo на человеке осложнено. Использование техники геномного мечения позволило нам визуализировать МСК в тканях мышей. Несмотря на то, что мышиные модели не всегда воспроизводятся на человеке, можно изучать поведение МСК у человека при развитии заболеваний с фиброзным компонентом. (человек или мышь) Отслеживание генетически меченых МСК позволило идентифицировать и охарактеризовать несколько тканевых популяций МСК. Показано, что пул МСК, находящихся в тканях включает в себя несколько популяций с различной мультипотентностью. Резидентные клетки могут включать периваскулярные мезенхимальные клетки предшественники и давать начало специфическим типам клеток соединительной ткани, а также интерстициальным фибробластам, необходимым для формирования базальной мембраны, окружающей паренхиматозные структуры. Несмотря на то, что связь между периваскулярными мезенхимальными клетками предшественницами и перицитами остается не совсем ясной [8], пМСК имеют особенности перицитов, которые анатомически определяют их как клетки стенки сосудов, частично

или полностью входящие в состав базальной мембраны капилляров [9]. Эти клетки контактируют с внутрисосудистой поверхностью эндотелия в микрососудистом русле всех тканей взрослого человека. Полученный органоспецифический паттерн МСК позволяет быстро обнаруживать

локальные сигналы из определенных тканей, включая сигналы при повреждении.

Стоит отметить, что данные критерии относятся только к клеткам человека. Так, например, у мышиных МСК профиль экспрессии поверхностных маркеров отличается. Кроме того, помимо перечисленных антигенов МСК могут экспрессировать широкий спектр самых разнообразных маркеров в зависимости от их текущего фенотипа. Среди прочего, для позитивного отбора МСК можно использовать маркеры: CD3, CD10, CD44, CD54, αSMA и другие [10].

Дальнейшее изучение МСК было продолжено в работах Каплана [11], который предложил идею мезенхимной стволовой клетки и объединил под этим названием фибробластоподобные клетки способные к адгезии к пластику с несколькими направлениями в дифференцировке в мезенхимные производные. Определяемая популяция включает в себя не только стволовые клетки, но и более зрелые, коммитированные к дифференцировке и утратившие один из важнейших атрибутов стволовых клеток - способность к самоподдержанию. Международное общество клеточной терапии рекомендовало называть популяцию таких клеток как «мультипотентные мезенхимные стромальные клетки», наравне используя термин «мезенхимные стволовые клетки» только для тех, которые удовлетворяют строгим критериям стволовости [12].

Известно, что популяция МСК неоднородна и отличается морфологическими, фенотипическими и физиологическими свойствами. Связано это, возможно со сложной организацией стромального дифферона и появляющаяся гетерогенность определяется во время закладки соединительной ткани в эмбриональном периоде.

В наблюдениях было отмечено, что способность клеток к пролиферации связана с неоднородностью клеток по размеру в работах по изучению клонального роста КОЕ-Ф [13]. Плотность популяций клеток также

неодинакова. Участки колоний, где пролиферация протекала активно, характеризовались большей плотность, в отличие от участков с меньшей плотностью, где пролиферация была выражена слабее [14]. Колониеобразующие единицы, дающие разряженные популяции обладали повышенной устойчивостью к воздействию ионизирующего излучения, в то время как плотные популяции обладали меньшей сопротивляемостью [15].

Неодинаковый пролиферативный потенциал разных клонов обнаруживается и при анализе пассируемых культур МСК [16]. Кроме того, в этих культурах описаны клетки нескольких морфологических типов — тонкие веретеновидные, крупные распластанные и мелкие округлые [17, 18, 19]. Различия в их морфологии отражают неодинаковые зрелость, пролиферативную активность и потенции к дифференцировке. Так, клонированная субпопуляция плоских клеток, выделенная из культуры костного мозга, в ходе пассирования быстро теряет адипо- и хондрогенные потенции, тогда как клонированные веретеновидные клетки сохраняют их [20]. По данным из литературы [21], веретеновидные клетки имеют более высокую скорость пролиферации по сравнению с плоскими (видимо, представляющими собой наиболее зрелую субпопуляцию МСК), а наибольшая скорость роста свойственна очень мелким круглым клеткам с высоким ядерно-плазменным отношением, названным RS-клетками (rapidly self-renewing cells). RS-клетки отличаются от остальных клеток той же культуры по фенотипу и имеют наибольшие потенции к дифференцировке.

Антигенный фенотип МСК также неоднороден. В то время как одни поверхностные маркеры (в частности, CD73, CD90 и CD105) стабильно экспрессируются большинством клеток этого типа, экспрессия других (например, Stro-1, CD106 и MSCA-1) варьирует, что, видимо, связано с различиями в потенциях и степени зрелости клеток [22]. Известно, что МСК костного мозга человека, несущие CD56, отличаются от лишенных его большей способностью к колониеобразованию, наличием хондрогенных и

отсутствием адипогенных потенций [23], а повышенный уровень экспрессии CD146 характерен для трипотентных клонов МСК в противоположность монопотентным [24]. Гетерогенность популяции МСК проявляется и в неодинаковой активности щелочной фосфатазы, различия в которой отмечаются как между клонами, образуемыми КОЕ-Ф [25, 26] так и в пределах одного клона [27]. Активность этого фермента в первичной культуре МСК не зависит от клоногенной способности клеток и присутствия на них антигенов CD105 и CD29, однако содержащие его клетки отличаются более крупным размером, меньшей скоростью роста и повышенными потенциями к остеогенезу [28].

Клетки в составе популяции МСК неодинаковы по чувствительности к цитотоксическим агентам. При введении животным бусульфана, метотрексата, циклофосфамида [29], 5-фторурацила [30] или дипина [31] часть КОЕ-Ф костного мозга выживает и сохраняет клоногенную способность. Во многих случаях чувствительность МСК к цитотоксическим препаратам коррелирует с их положением в гистогенетическом ряду. Так, к ингибитору синтеза ДНК цитозинарабинозиду устойчивы клетки с высоким пролиферативным потенциалом, образующие очень крупные колонии [32]. При обработке 5-фторурацилом, по некоторым данным, избирательно сохраняются покоящиеся некоммитированные клетки, способные к самоподдержанию[33], тогда как алкилирующий препарат дипин поражает, по-видимому, наиболее молодую категорию МСК с высоким репаративным потенциалом [34].

Неодинаковы и адгезивные свойства МСК. При посеве в первичную культуру часть КОЕ-Ф прикрепляется к субстрату в первые часы или дни, тогда как другие клетки длительное время остаются во взвеси, сохраняя клоногенность [35, 36]. Математический анализ структуры популяции стромальных клеток костного мозга выявил в ней субпопуляции с различной степенью адгезии к пластику и фибронектину [37]. Взаимосвязь адгезивных

свойств МСК с другими их характеристиками не вполне ясна. Есть данные о том, что низкая адгезивность к пластику свойственна наиболее ранним стромальным клеткам [38]; с другой стороны, сравнение фракций МСК, прикрепляющихся к пластику в разные сроки, не обнаруживает различий в их чувствительности к факторам роста и способности к основным дифференцировкам [39, 40].

Наконец, клоны и субпопуляции МСК, даже полученные из одного источника, различаются широтой спектра потенций к дифференцировке [41, 42, 43, 44] и их выраженностью [45, 46, 47].

Отчасти различия в свойствах МСК связаны с влиянием микроокружения. Так, клетки из центральных и периферических областей одной и той же клональной колонии могут иметь разные остеогенные и адипогенные потенции. Предполагаемая причина этих различий — неодинаковая плотность расположения клеток в различных участках колонии, влекущая за собой различия в экспрессии регуляторных молекул (в частности, ингибитора сигнального пути Wnt Dkk-1) и компонентов внеклеточного матрикса, что предрасполагает клетки к тому или иному направлению дифференцировки [48, 49]. Но, несомненно, в значительной степени гетерогенность морфологических, фенотипических и функциональных характеристик МСК отражает внутренние различия между клетками, занимающими то или иное положение в гистогенетическом ряду, и может свидетельствовать о сложной, к настоящему времени еще недостаточно изученной иерархической структуре популяции.

Таким образом, видно, что популяция МСК объединяет клетки, различные по маркерам и функциям, которые могут играть роль в физиологических процессах.

1.2 Роль МСК в фиброзе.

Фиброз органов и связанная с ним недостаточность вносит большой вклад в смертность во всем мире. Некоторые фибротические заболевания являются идиопатическими, в то время как другие имеют хорошо описанную этиологию, связанную с образом жизни, генетической предрасположенностью и другими системными нарушениями и осложнениями [50].

Большое количество дегенеративных расстройств, связанных с фиброзом, остаются заболеваниями с высокой заболеваемостью и смертностью из-за медленного развития методов лечения этого нарушения. Частично это связано с недостаточным пониманием биологии участия МСК в развития фиброза и механизмов регуляции фибротических процессов в тканях.

Представление о том, что фиброз в разных тканях обусловлен практически идентичной популяцией МСК и схожим набором внешних сигналов постепенно устаревает. Анатомическое распределение и организация резидентных популяций МСК в разных органах определяются во время эмбрионального развития, направляется уникальной комбинацией внешних сигналов, что в конечном итоге формирует совершенно определенную тканеспецифичность. Мнение о том, что специфичность МСК определяет происхождение и локальное микроокружение подтверждается экспериментами по отслеживанию происхождения МСК на мышах. В подобных экспериментах было показано, что направления развития клеток из разных источников эмбрионального происхождения впоследствии заселяют анатомически схожие места в различное время развития. Эти направления начинаются из двух источников – нейроэктодермы и первичной полоски (через мезодерму), дающие начало соединительной ткани во всем организме [51, 52]. Несмотря на фенотипическое сходство, производные от популяций этих клеток предшественниц в тканях взрослого организма обладают

уникальными эпигенетическими свойствами и сохраняют присущие им свойства и направления развития [53, 54 ,55].

Последние достижения в фенотипировании, выделении, манипуляциях с клетками, визуализации, трансплантации и отслеживании позволили идентифицировать, характеризовать и отследить судьбу клеток в нормальном состоянии и при нарушениях. Отслеживание с помощью генетических меток является золотым стандартом для изучения клеточного происхождения миофибробластов при фиброзе.

Основным типом клеток, общим для многих известных фибротичексих заболеваний, являются миофибробласты. Миофибробласты это ACTA2+ клетки, которые могут откладывать белки внеклеточного матрикса, такие как коллаген и фибронектин. Фиброз характеризуется прогрессирующим накоплением большого количества миофибробластов. Это приводит к чрезмерному отложению миофибробластами компонентов внеклеточного матрикса, который нарушает структуру и функцию органа и в конечном итоге может привести к смерти. В таких условиях поражённые органы не способны выполнять свою функцию [56].

Идентификация клеток-предшественниц из миофиброластов имеет важное клиническое значение, поскольку открывает путь к лучшему пониманию патологических процессов, которые вызывают фиброз на клеточном уровне и разработке селективных антифибротических препаратов, направленных на определенные типы клеток или передачу сигналов, управляющих фиброзом. В настоящее время не существует антифибротических препаратов, которые могли бы полностью остановить прогрессирование фибротических изменений, а доступная терапия основана на ингибировании рецепторных тирозинкиназ, которые играют решающую роль в передаче клеточных сигналов в здоровых клетках и тканях.

Считается, что популяция миофибробластов гетерогенна и происходит из перицитов, резидентных фибробластов и клеток костного мозга. Спорным

моментом является утверждение о том, могут ли эндотелиальные клетки вносить вклад в пул миофибробластов [57, 58, 59]. TGFb рассматривается как главный регулятор дифференцировки миофибробластов при фиброзе, что доказывается несколькими исследованиями [60, 61, 62, 63, 64].

Еще одной концепцией развития фиброза является пластичность фибробластов, при которой адипоциты или адипоцитоподобные клетки превращаются в коллагенпродуцирующие миофибробласты [65]. Обратный путь наблюдали во время разрешения фиброза, когда миофибробласты, образовавшиеся во время фиброза возвращаются обратно в адипоцитоподобные клетки и адипоциты [65].

Судьба ассоциированных с фиброзом миофибробластов была изучена недостаточно после разрешения фиброза на некоторых животных моделях. Изучение судьбы миофиброластов после разрешения фиброза на животных моделях не менее важно для установления исходных клеток для определения путей, которые критичны для разрешения фибротических изменений.

Колониеобразующие единицы-фибробласты, также называемые мультипотентными мезенхимальными стромальными клетками или просто МСК, были впервые описаны как прикрепленные к пластику фибробластоподобные клетки, которые образуют колонии in vitro [66].

Внеклеточные везикулы играют важную аутокринную/паракринную роль в межклеточной коммуникации. Во внеклеточных везикулах находятся белки, мРНК, микроРНК, которые могут принимать участие в передаче информации клеткам-реципиентам в различные органы [67]. В зависимости от происхождения внеклеточные везикулы выполняют различные функции [68]. ВВ из МСК обладают терапевтической активностью, которую можно сравнить с самими МСК. Недавние исследования на животных моделях показывают, что ВВ, происходящие из МСК, обладают значительным потенциалом в качестве новой альтернативы клеточной терапии. По сравнению с самими МСК, ВВ имеют более высокий профиль безопасности и

могут храниться без потери функции. Было отмечено, что ВВ, полученные из МСК, подавляют провоспалительные процессы и окислительный стресс, фиброз тканей и ремоделирование в различных моделях воспалительных заболеваний in vivo путем передачи их компонентов. Однако существуют серьезные ограничения по внедрению этой технологии в клинику [69, 70 ,71].

1.3 Роль нкРНК в фиброзе

Экспрессия генов является способом реализации генетической информации клетки через транскрипцию её в матричные (кодирующие) РНК, и затем трансляции с них посредством рибосом в аминокислотные последовательности. Данный процесс строго регулируется на каждом этапе бесчисленным множеством различных механизмов. Все не кодирующие белок РНК, в том числе рРНК и тРНК, объединяют под общим названием - некодирующие РНК (нкРНК). В клетках эукариот существует особый подкласс нкРНК, отвечающий за регуляцию экспрессии генов – микроРНК [72].

МикроРНК представляют из себя короткие (21-23 нт) нкРНК, отвечающие за снижение экспрессии, или сайленсинг, генов путём взаимодействия с соответствующими мРНК, сайтами активации транскрипции и белками. МикроРНК отвечают за регуляцию экспрессии около 60% всех генов человека. Известно около 1800 предшественников микроРНК, дающих начало более чем 2500 зрелых молекул [73].

Для микроРНК существует собственная общепринятая номенклатура. Для присвоения идентификационного номера необходимо экспериментальное подтверждение существования данной микроРНК. Каждый идентификатор имеет приставку “mir”, регистр букв несёт следующую информацию:

- mir - обозначение пре-микроРНК;

- MIR - обозначение гена, кодирующего микроРНК;
- miR - обозначение зрелой формы микроРНК.

Перед приставкой может добавляться трёхбуквенный код, обозначающий вид, из которого была выделена данная микроРНК. Например, приставка “hsa” означает *Homo sapiens.* Вирусные микроРНК обозначаются приставкой “v”, микроРНК *Drosophila melanogaster* - приставкой “d”. После приставки “mir” через дефис следует порядковый номер микроРНК, соответствующий порядку её публикации. Две микроРНК, представляющие из себя практически идентичные последовательности, отличающиеся на один или два нуклеотида, имеют одинаковый порядковый номер, к которому без дополнительных разделителей приписывается строчная буква (например, hsa-mir-156a). В случае, когда гены, дающие начало полностью идентичным зрелым микроРНК, находятся в разных частях генома, к идентификаторам соответствующих генов и пре-микроРНК после порядкового номера через дефис добавляется ещё одна цифра. Если две зрелых микроРНК получаются из общей пре-микроРНК, но с разных её концов (изомикроРНК, изомиРы), к идентификатору добавляется ещё один суффикс, обозначающий, с какого именно плеча шпильки образовалась данная микроРНК: 3p - 3’-конец, 5p - 5’-конец. Кроме того, можно встретить суффиксы “s” и “as”, обозначающие смысловое и антисмысловое происхождение микроРНК соответственно, по сути, означающие то же, что и 3p и 5p. Если известен уровень экспрессии для каждого из изомиРов, наименее представленная микроРНК обозначается символом *. Для некоторых микроРНК, открытых ещё до создания единой номенклатуры, были сохранены их исторические названия, чаще всего отражающие мутантную линию, из которой они были получены. Это относится, например, к микроРНК семейств let-7 и lin-4, открытых при изучении линий *Caenorhabditis elegans,* нокаутным по генам с соответствующими названиями [74, 75, 76].

Существует несколько путей биогенеза микроРНК (Рис. 1). Большинство микроРНК имеют свой собственный промотер и располагаются кластерами, напоминающими полицистронные единицы, которые транскрибируются при классическом пути. Транскрипцию осуществляет РНК-полимераза II. Полученный полиаденилированный продукт называется при-микроРНК. За счёт наличия в последовательности комплементарных участков, при-микроРНК имеет структуру петли, содержащей неспаренные основания (мисмэтчи) и некомплементарные участки (выпетливания). Данный транскрипт процессируется в ядре комплексом DROSHA (или RNASEN) либо DGCR8, в результате чего при-микроРНК разрезается на более короткие двуцепочечные РНК, называемые пре-микроРНК. Далее при помощи белка экспортин 5 пре-микроРНК доставляются в цитоплазму, где комплекс DICER, содержащий домен белка-аргонавта, при взаимодействии с белком TRBP отрезает запетленную часть от пре-микроРНК с образованием двухцепочечного предшественника микроРНК. Важной особенностью является наличие 2х- или 3х-нуклеотидных довесков с обеих концов дцРНК. Иногда для процессинга пре-микроРНК DICER не требуется, а достаточно только белка-аргонавта. На последней стадии вокруг белка DICER, связанного с дуплексом, собирается комплекс RISC. Происходит отбор одной из цепей дуплекса и её встраивание в RISC. Комплекс со встроенной зрелой микроРНК далее опознаёт целевую, или таргетную, мРНК [77].

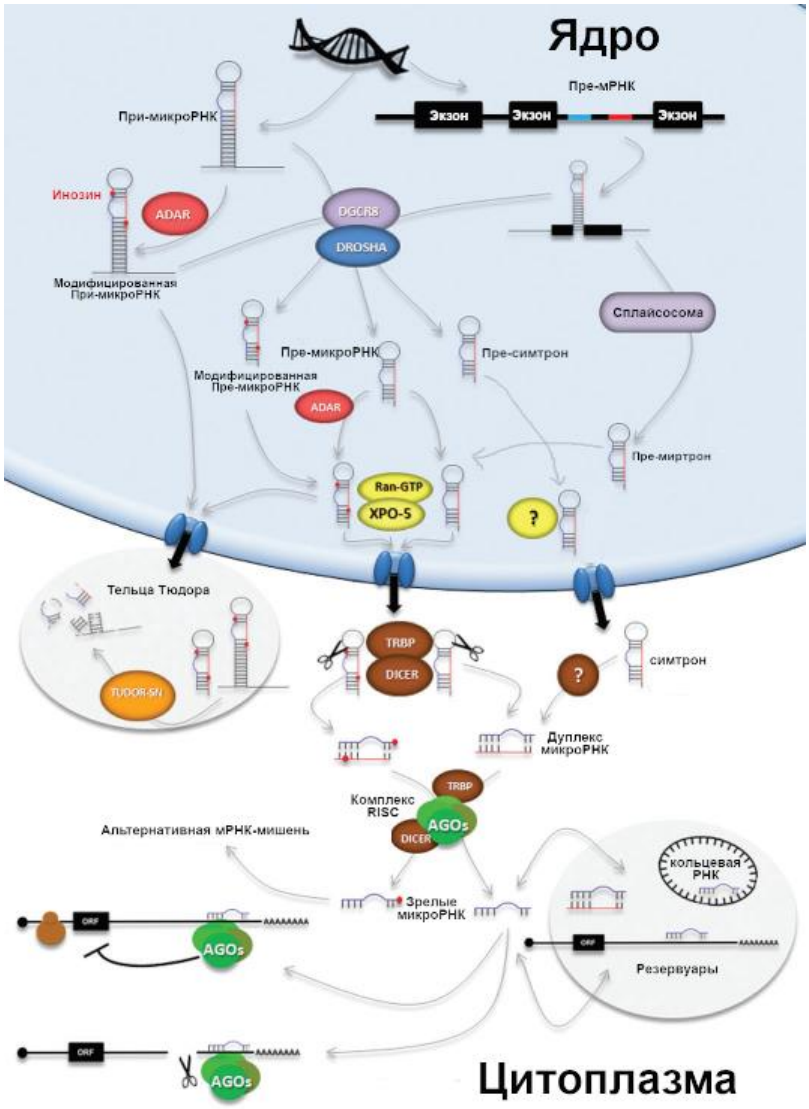


Рисунок 1**.** Биогенез, функция и утилизация микроРНК [78].

При альтернативных путях синтеза микроРНК имеют несколько иное генетическое происходение (Рис. 2). Их последовательности закодированы в интрагенной области, чаще всего в интронах, однако иногда могут включать в себя и экзонную часть. МикроРНК, берущие начало в интронном участке мРНК, называются миртронами. Миртроны образуются в независимый продукт, называемый пре-миртроном, после процессинга мРНК сплайсосомой. Они имеют структуру, напоминающую пре-микроРНК. После образования они доставляются из ядра в цитоплазму белком экспортин 5, где далее идут по классическому пути. Таким образом, данный альтернативный путь можно назвать DROSHA/DGCR8 независимым. Стоит отметить, что некоторые пре-миртроны имеют на одной из цепей довесок, который должен

быть убран экзонуклеазой перед тем, как DICER приступит к работе. Другим альтернативным источником микроРНК являются так называемые симтроны. Подобно миртронам, симтроны происходят из мРНК, однако не в результате действия сплайсосомы, а с помощью белка DROSHA. Интересной особенностью симтронов является то, что для их транспортировки в цитоплазму не требуется белок экспортин 5, а процессинг до зрелой формы микроРНК является независимым от комплексов DICER и белков-аргонавтов [79, 80].

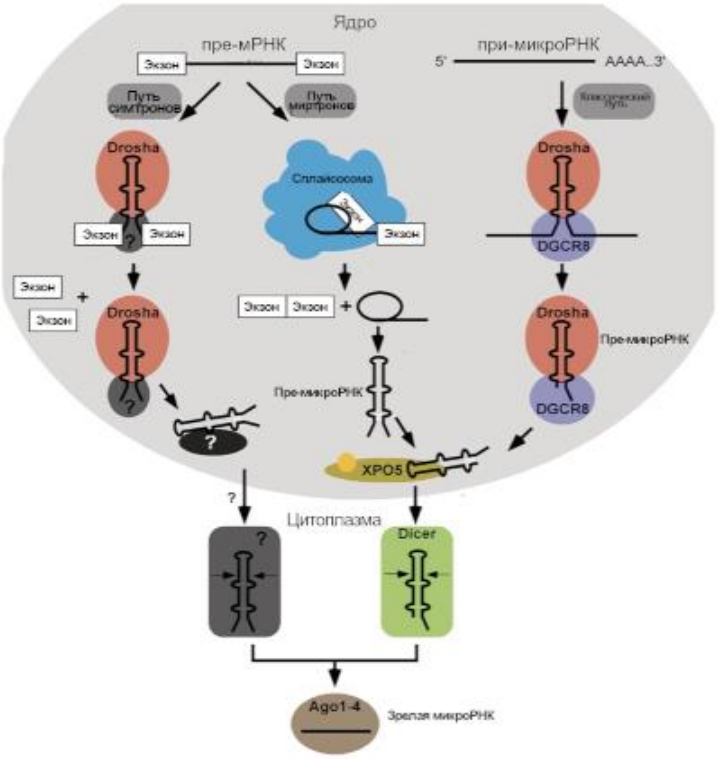


Рисунок 2**.** Альтернативные пути биогенеза микроРНК [81].

МикроРНК содержат в себе последовательность из 6-8 нуклеотидов, которая отвечает за узнавание таргетной мРНК. Как правило, связывание происходит в 3’-некодирующей области мРНК (3’-UTR). Ключевую роль в связывании с матричной РНК играют белки-аргонавты, входящие в состав комплекса RISC. Они ориентируют микроРНК для её взаимодействия с таргетной мРНК. В зависимости от класса белка-аргонавта, далее может происходить либо разрезание таргетной мРНК, либо недопущение дальнейшей её трансляции за счёт её связывания. На выбор судьбы мРНК

может влиять не только состав комплекса RISC, так и степень сродства микроРНК к своей мишени. Дело в том, что для связывания микроРНК не обязательно должна быть строго комплементарна своей таргетной последовательности. Чем выше сродство, тем больше вероятность деградации мРНК. Одна микроРНК может связываться с несколькими матричными РНК, также как и несколько разных микроРНК могут таргетировать одну и ту же мРНК, часто в таких случаях проявляя синергический эффект на подавление экспрессии гена [82].

Пре-микроРНК содержит в себе два плеча, с 3’ и 5’ концов, оба из которых потенциально могут стать микроРНК. Две эти последовательности не полностью комплементарны. Каким образом происходит выбор цепи для встраивания остаётся неясным. Предполагается, что это может быть связано с термодинамическими характеристиками комплекса микроРНК-RISC, либо наличием вторичных структур [83, 84]. Раньше считалось, что в большинстве случаев только одна цепь из пре-микроРНК функциональна, другая же деградируется. Однако согласно последним исследованиям, это не так. Более того, количество микроРНК одного и другого плеча может отражать функциональное состояние клетки. Так, известно, что наиболее представленной является 3’-форма микроРНК, однако последние исследования показывают, что с возрастом происходит сдвиг в сторону увеличения количества микроРНК 5’-плеча [85].

Разнообразие регулирующих свойств микроРНК повышается также за счёт внесения в них корректировок. Так, например, белки семейства ADAR могут заменять аденозин на инозин, что может влиять на взаимодействие с дуплекса с DICER и даже на узнавание таргетной мРНК. Кроме того, отрезание петлевой части белком DICER с образованием дуплекса может происходить со сдвигом сайта внесения разреза, что приводит к изменению последовательности зрелых микроРНК. Такие микроРНК, отличающиеся в зрелом состоянии от своего предшественника, называются изомиры [86, 87].

Время полужизни микроРНК в клетках достаточно велико и может составлять до нескольких суток. Одна молекула микроРНК может взаимодействовать с несколькими молекулами мРНК, вплоть до ста. Регуляция транскрипции микроРНК может происходить самими микроРНК, например, через связывание с их собственной мРНК, либо через связывание с микроРНК соответствующими транскрипционными факторами. Предполагается также возможность существования в клетках резервуаров для микроРНК. Среди прочего роль таких резервуаров могут играть кольцевые РНК через связывание комплементарных микроРНК [88].

Исходной функцией РНК-интерференции и, в частности, микроРНК была защита от ретровирусных транспозонов. Вероятно, в процессе эволюции эта система постепенно была рекрутирована для использования в регуляции генной экспрессии [89]. Влияние микроРНК на экспрессию генов нельзя назвать бинарным - “включить - выключить”. МикроРНК влияют на это более тонко, изменяя уровень экспрессии гена. В первую очередь на это влияет содержание данной микроРНК в клетке, которое может отличаться на разных этапах развития клетки, а также в разных тканях. Интересно, что не всегда микроРНК вызывают снижение экспрессии гена. Так, miR-373 комплементарна последовательности промотера белка Е-кадгерина. Связывание с этим участком микроРНК привлекает РНК-полимеразу II, что повышает транскрипцию этого гена. Важной особенность микроРНК является выполнение ими противоположных функций в зависимости от клеточного контекста. Так, например, некоторые микроРНК широко представлены в одних видах злокачественных образований, однако их экспрессия подавлена в других, поскольку в первом случае они проявляют онкогенные свойства, а в других - онкосупрессивные. Профили экспрессии различных микроРНК сильно варьируют между здоровыми и больными тканями, организмами. В связи с этим большой интерес представляет изучение представленности маркерных микроРНК в межклеточной среде и биопсиях [90].

Как важный регуляторный элемент, микроРНК широко представлены в клеточных экзосомах и, в частности, в экзосомах МСК. Изучение микроРНКового содержания везикул представлят большой интерес для регенеративной медицины, поскольку с помощью микроРНК МСК могут формировать ответ на различные стимулы и регулировать развитие таких процессов, как фиброз.

1.4 Типирование и гетерогенность

Идея типирования клеток в результатах обработки данных scRNA-seq возникла после попытки объединения данных о дифференциально экспрессирующихся генах в разных субпопуляциях и имеющейся информации о специфических маркерах для большого количества клеток. Проблема и научный вызов в этой идее состоит в том, что методы автоматического типирования не учитывают переходных, промежуточных форм клеток, в которых они находятся в процессе дифференцировки. Поскольку scRNA-seq представляет собой "снимок" клеток в моменте, необходимо понимать, что в данный момент не все клетки обязательно находятся в конечно дифференцированном состоянии, большинство из них находятся в промежуточных состояниях, которые можно охарактеризовать через преобладание биологических процессов, посредством которых клетки стремится достичь целевого состояния. На сегодняшний день не существует программ, объединяющих в себе типирование и по специфическим маркерам, и по биологическим процессам, что крайне важно для правильной интерпретации результатов.

Типирование клеток - трудная научная задача, которая требует системного подхода. На сегодняшний день эта задача решена лишь частично благодаря большому количеству фактических и экспериментальных данных. На сегодняшний день существует большое количество клеточных референсов, содержащих информацию о специфических маркерах многих типов клеток. Но известно, что при выделении РНК из образца получается

лишь «снимок» транскриптома клеток в момент времени и нельзя быть уверены в том, что все клетки образца находятся в своих конечно дифференцированных состояниях, а следовательно, они не могут быть идентифицированы по своим специфическим маркерам.

Частично вопрос может быть решен анализом биологических процессов, происходящих в клетках образца на момент секвенирования. Это может помочь понять, какие процессы для клетки актуальны в текущий момент и частично ответит на вопрос о направлении развития клетки. Но такой подход не заменяет метода типирования клеток по специфическим маркерам, а лишь дополняет его.

При вычислении векторов RNA-velocity для клеток образца, вдоль образующихся из отдельных векторов траекторий развития могут находиться кластеры клеток с неизвестным типом. Знания о типе исходной и конечной популяции клеток дает нам возможность с высокой вероятностью предположить тип клеток в кластере, находящемся между этими популяциями. Но и такой подход недостаточно точно предсказывает тип клеток.

Фундаментальные знания о дифферонах, промежуточных формах клеток на пути их дифференцировки и накопленные данные о транскриптоме клеток в этих переходных формах могут значительно продвинуть научное сообщество в решении проблемы типирования клеток не в конечно дифференцированном состоянии.

Последние достижения в технологии секвенирования единичных клеток позволяют получать профиль экспрессии в масштабе всего генома в отдельных клетках с высокой производительностью. Для определения «типов клеток», необходимы значительные усилия по правильному анализу результатов секвенирования единичных клеток, которые образуют составные части, элементы сложной ткани. Классификация типов клеток по данным секвенирования единичных клеток включает применение вычислительных

инструментов, основанных на уменьшении размерности и кластеризации, статистическому анализу для определения молекулярных сигнатур, уникальных для каждого типа. Поскольку количество данных продолжают расти, существует множество вычислительных задач, требующих масштабируемости аналитических методов, гибкости и надежности. Кроме того, необходимо внимательно рассмотреть экспериментальные ошибки и статистические проблемы, которые являются уникальными для этих измерений для того, чтобы избежать артефактов [91].

Современные методы определения типов клеток обычно включают использование неконтролируемой кластеризации, идентификацию сигнатурных генов в каждом кластере с последующим поиском этих генов вручную в литературе и базах данных для определения типов клеток. Однако есть несколько ограничений, связанных с этими подходами, такими как нежелательные источники вариаций, которые влияют на кластеризацию и отсутствие канонических маркеров для определенных типов клеток. Существует метод автоматического типирования клеток с использованием нейронных сетей ACTINN, в котором используется нейронная сеть с тремя скрытыми слоями, обучающаяся на наборах данных с заранее заданными типами клеток и прогнозирует типы клеток для других наборов данных на основе обученных параметров [92].

Для идентификации типов клеток из данных scRNAseq с высокой производительностью могут быть также задействованы нейронные сети. Однако для этого требуется большое количество единичны клеток с точно аннотированными типами для построения идентифицируемой модели. К сожалению, определение типов клеток в данных scRNAseq неэффективный и трудоемкий процесс, так как требует постоянного контроля результатов в процессе специфических маркерных генов. Для решения этой задачи существует модель полу-контролируемого обучения для использования немеченых клеток scRNAseq и ограниченное количество меченых клеток

scRNAseq для реализации идентификация клеток. С помощью такой модели реализуется полууправляемый метод обучения на основе рекуррентных сверточных нейронных сетей (RCNN), который включает в себя общую сеть, контролируемую сеть и неконтролируемая сеть. Предлагаемая модель оценена на macosko2015 - крупномасштабный набор транскриптомных данных по отдельным клеткам с достоверно определенными отдельными типами клеток. Замечено, что предлагаемая модель способна достичь обнадеживающих характеристик обучаясь на очень ограниченном количестве меченых клеток scRNAseq вместе с большим количеством немеченых клеток scRNAseq [93].

Одноклеточная геномика и протеомика позволяют не только изучить состояние клетки, но также обеспечивает высокое разрешение переходов между состояниями. Эти измерения могут, наконец, объяснить метафору, что Ч. Уоддингтон озвучивал почти 60 лет назад, чтобы объяснить клеточную пластичность: клетки являются резидентами обширного «ландшафта» возможных состояний, по которому они путешествуют во время развития и при патологии. Одноклеточная технология помогает не только находить клетки на этом ландшафте, но и освещает молекулярные механизмы, которые формируют сам пейзаж. Однако одноклеточная геномика - это область, находящаяся в зачаточном состоянии, для полной реализации потенциала которой необходимо развитие экспериментальных и вычислительных технологий [94].

В отличие от массовых измерений, которые усредняют экспрессию генов по индивидууму клеток, измерения генов в отдельных клетках можно использовать для изучения нескольких различных тканей и органов на разных этапах развития. Определение типов клеток, присутствующих в образце, по данным транскриптома отдельных клеток является общей целью многих одноклеточных экспериментов. Для этого было разработано несколько методов. Однако правильное определение истинных типов клеток

остается проблемой. Гипотеза определения типов заключается в том, что значимые характеристики данных остаются, несмотря на небольшие искажения данных. Эффективность предложенного метода была проверена на восьми общедоступных наборах данных scRNA-seq с известными типами клеток, а также пяти наборах данных смоделированных с различной степенью кластеризации. Предложенный метод сравнивали с пятью другими существующими методы: RaceID, SNN-Cliq, SINCERA, SEURAT и SC3. Результаты показывают, что предлагаемый метод работает лучше, чем существующие методы [95].

Сложные ткани, такие как мозг, состоят из нескольких различных типов клеток, каждый из которых имеет различные и важные роли, например, в нервной функции. Кроме того, недавно было установлено, что клетки, составляющие эту субпопуляцию, сами типы обладают значительной межклеточной гетерогенностью, в частности, на уровне экспрессии генов. Способность к исследованию этой неоднородности было революционным благодаря достижениям в экспериментальных технологиях, таких как Whole mount in situ Гибридизации (WiSH) и секвенирование одноклеточной РНК. Следовательно, теперь можно изучать уровни экспрессии генов в тысячах клеток из одного и того же типа ткани. После создания таких данных одной из ключевых целей является кластеризация клеток в группы, которые соответствуют как известным, так и предположительно новым типам клеток. Хотя существует множество алгоритмов кластеризации, они обычно не в состоянии включить информацию о пространственной зависимости между клетками в исследуемой ткани. Когда такая информация существует, она дает важную информацию, которая должна быть непосредственно включена в схему кластеризации. Был разработан метод кластеризации, который использует модель скрытого марковского случайного поля (HMRF) для использования количественных мер выражения и пространственной информации. Чтобы точно отразить основную биологию, были расширены текущие HMRF подходы, позволяя степени пространственной когерентности

различаться между кластерами. Использование смоделированных данных перед его применением для кластеризации данных экспрессии генов отдельных клеток, сгенерированных путем применения WiSH позволяет изучить паттерны экспрессии. Подход позволяет идентифицировать отдельные типы клеток, а также выявить новые, ранее неизученные типы клеток в мозге этой важной модельной системы [96].

Итеративно применяя подход машинного обучения к заданному набору клеток, можно идентифицировать отдельные группы клеток и взвешенный список генов признаков для каждой группы. Дифференциально выраженные гены признаков отличают данную группу клеток от других клеток. Каждая такая группа клеток соответствует предполагаемому типу клеток или состоянию, характеризующееся признаками генов как маркеров. Сравнительный анализ с использованием аннотированных экспертами наборов данных scRNA-seq показывает, что метод автоматически идентифицирует типы клеток с высокой точностью [97].

1.5 Современные подходы к анализу транскриптома одиночных клеток

С момента появления секвенирования накопилось большое количество транскриптомных данных. Несмотря на активное изучение этих данных и их углубленный анализ до сих пор невозможно ответить на многие вопросы, касающиеся науки и медицины.

Долгое время количество транскриптов отдельных генов оценивалось в образце в целом, без разделения на отдельные клетки. К сожалению, такой подход дает возможность оценить лишь общую реакцию изучаемой ткани на предъявляемое воздействие. Однако, давно известно, что в популяции клеток не все клетки реагируют на один и тот же стимул одинаково. Этот факт постепенно приближал научное сообщество к созданию метода, который мог бы оценивать изменение транскрипционного профиля каждой отдельной клетки.

Естественно, новые идеи запустили процесс изобретения новых технологий и платформ, сложность устройства которых заметно возросло, а также потребовало разработки новых наборов реагентов для изучения образцов на уровне отдельных клеток.

В первых появившихся платформах была возможность изучать несколько сотен клеток. Такого количества клеток оказалось недостаточно для изучения гетерогенности, а также поиска редких типов клеток, которые невозможно было идентифицировать среди небольшого количества.

Позже стали появляться платформы, где количество анализируемых клеток возросло до десятков тысяч. Одновременно с увеличением количества анализируемых клеток усложнялись протоколы пробоподготовки и биоинформатического анализа получаемых данных и, как следствие увеличивались технические требования к компьютерам, на которых производятся подобные вычисления.

Для изучения гетерогенности ответа клеток необходимо использовать современные методы, благодаря которым можно исследовать транскриптом единичных клеток. Метод 10x Genomics появился не так давно и основан на пришивании UMI к каждому транскрипту в каждой клетке, и баркодов, по которым можно определить принадлежность транскрипта к определенной клетке.

Таким образом, появилась возможность изучать транскриптомный профиль и уровень экспрессии каждой отдельно взятой клетки и после кластеризации увидеть результат группировки клеток по схожести паттернов экспрессии. Важным требованием пробоподготовки является получение образца клеток. Основная задача пробоподготовки – получение суспензии жизнеспособных неагрегированных клеток (диссоциация клеток). Необходимо определиться с концентрацией клеток, так как для удачного захвата требуется определенное их число в зависимости от выбранного метода изоляции клеток: от десяти клеток (при использовании

микропипетирования, цитоплазматической аспирации, лазерной микродиссекции) до тысяч клеток (при использовании приборов на основе технологий FACS, микрофлюидики и микрокапель). В случае работы с высокоэффективными приборами, например Chromium 10X Genomics, концентрация клеток должна быть примерно 106 клеток в мл. При работе с животными или тканями, содержащими число клеток меньше необходимого, нужно увеличивать число образцов на пробу. Все этапы диссоциации клеток проводят в минимальном объеме раствора (от 50 до 1000 мкл) для повышения концентрации и уменьшения возможных потерь клеток.

Протокол секвенирования транскриптомов одиночных клеток состоит из трех этапов: обратная транскрипция, амплификация кДНК (WTA – полнотранскриптомная амплификация) и подготовка библиотеки. Несмотря на нежелательность амплификации кДНК (из-за возможности возникновения ошибок полимеразы или потери редких транскриптов), данный этап необходим для создания библиотеки, так как принято, что количество общей РНК в клетке составляет около 10 пг, что недостаточно для успешного секвенирования. В зависимости от задачи и используемой платформы для изоляции клеток будут отличаться протоколы для обратной транскрипции и получения кДНК.

В настоящее время можно выделить три основных подхода.

Первым был предложен метод с использованием олиго-dT-праймеров, конъюгированных с адаптерами, для обратной транскрипции и избирательной амплификации полиаденилированной мРНК с помощью ПЦР. Этот протокол имеет существенный недостаток: из-за смещения в область сгенерированных 3′-концов во время обратной транскрипции происходит потеря информации для анализа альтернативного сплайсинга.

Позже был разработан подход, позволяющий конструировать полноразмерную кДНК, – так называемый синтез кДНК со сменой матрицы (template switching cDNA synthesis). Преимущество данного метода

заключается в получении и амплификации полноразмерной кДНК, что позволяет определять варианты альтернативного сплайсинга и аллель-специфическую экспрессию (ASE). Такой подход используется в протоколах STRT, SMART-seq и SMART-seq2. Особенность перечисленных выше протоколов – амплификация, в ходе которой происходит экспоненциальный рост числа транскриптов, что будет приводить к смещению в ходе анализа и потере минорных экспрессированных генов. В качестве альтернативы был разработан подход транскрипции in vitro (IVT) для линейной амплификации кДНК, который представлен в таких протоколах для анализа единичных клеток, как CEL-Seq и MARS-Seq.

Третий подход заключается в дополнительном использовании уникальных молекулярных идентификаторов (UMI), представляющих собой случайные короткие последовательности от 6 до 10 п. н., встраиваемые в олиго-dT-праймер и помогающие различить отдельные молекулы. Эта технология показана в таких протоколах для scRNA- seq, как CEL-Seq и CELSeq2, Drop-seq, MARS-Seq, SCRB-seq, STRT, In-Drop. Один из последних протоколов с использованием молекулярных идентификаторов – Quartz-Seq2, позволяет анализировать до 1536 клеток из одной пробы и повышает эффективность преобразования UMI с 22 % (для других протоколов scRNA-seq) до 35 %. Это дает возможность получить информацию о большом числе генов.

Последние достижения в параллельной работе с тысячами клеток потребовали усовершенствования баркодирования транскриптов. Наиболее современным и инновационным подходом, используемым в платформах, основанных на микрофлюидике и технологии микрокапель, служит применение дополнительно клеточного баркода (олигонуклеотид длиной ~14 п. н.) одновременно с праймерами, несущими на себе UMI, которые потом помещаются в каждую каплю с отдельными клетками. Клеточный баркод служит идентификатором всех последовательностей нуклеотидов из

различных клеток. Преимущество такого двойного баркодирования – высокая точность и возможность определения клетки, из которой получена каждая отдельная РНК. Для секвенирования на платформе Drop-Seq разработан протокол STAMPs (Single-cell Transcriptomes Attached to Microparticles) и протокол Cell-Seq – для платформы.

Наиболее высокопроизводительная коммерческая платформа Chromium 10X Genomics интегрировала технологию Gemcode, которая разделяет в каплях длинные молекулы ДНК и баркодирует их для создания библиотек под секвенирование. Использование двух баркодов на Chromium 10X Genomics при работе с единичными клетками позволяет уменьшить технический шум и проанализировать одновременно тысячи различных клеток, идентифицируя принадлежность каждого транскрипта, что особенно актуально при работе со сложными тканями. Это дает возможность определять профили экспрессии генов в масштабе одной клетки.

Существуют различные адаптеры, позволяющие подготовить библиотеки микроРНК, например 3′-концевой адаптер, содержащий 5′,5′-аденил пирофосфорилированный участок.

Секвенирование одиночных клеток и одноядерное секвенирование сделало возможным изучать клетку на молекулярном уровне. За последние годы количество методик секвенирования и вычислительных методов анализа данных значительно увеличилось. Стандартный протокол исследования транскриптома единичных клеток включает в себя пробоподготовку, выделение клеток, подготовку библиотек, обработку данных секвенирования и сырых данных, а также визуализацию и дополнительные методы обработки данных. Для подготовки суспензии единичных клеток существует множество протоколов, так как для разных тканей есть свои особенности выделения и подготовки клеток.

Для платформ с лунками клетки обычно переносят в планшеты с микро- или нанолуночками с использованием методов пипетки или лазерного

захвата, таких как сортировка флуоресцентно-активированных клеток (FACS) на основе поверхностных маркеров. Эта опция делает хорошо ориентированные платформы особенно полезными, когда требуется изоляция определенного подмножества клеток, например, для исследования редких типов клеток. Другим преимуществом является возможность визуального осмотра захваченных клеток, что позволяет идентифицировать лунки, содержащие поврежденные клетки или не содержащие их, и / или предоставляя дополнительную морфологическую информацию. Главный недостаток хорошо основанных платформ заключается в том, что они часто имеют низкую пропускную способность и требуют значительного количества ручной работы на клетку в отличие от других методов. Эти недостатки в некоторой степени преодолеваются за счет использования микрожидкостных платформ, таких как Fluidigm C1 (13), которые могут быть интегрированы в рабочий процесс некоторых платформ на основе микролунок, обеспечивая более высокую пропускную способность. Однако только около 10% клеток обычно захватываются микрофлюидной платформой, что делает ее непригодной для обнаружения редких типов клеток. Система C1 также позволяет проводить визуальный осмотр под микроскопом, тем самым позволяя пользователю исключить пустые лунки и лунки, содержащие поврежденные клетки или дублеты, до последующего приготовления библиотеки. Высокая стоимость микрожидкостных картриджей может ограничить размер образца, используемого в каждом проекте, но можно снизить расходы на реагенты, поскольку реакции можно проводить в меньшем объеме.

В методах на основе капель используется микрофлюидика, каждая отдельная клетка оказывается вместе с шариком внутри капли, которая включает определенные ферменты, необходимые для создания библиотеки. Шарик несет праймеры с уникальным баркодом, который связывает мРНК клетки и прикрепляются ко всем транскриптам клетки. Все капли можно объединить для создания библиотеки секвенирования. После секвенирования

транскриптов можно назначить исходной клетке на основе баркодов. Затраты на подготовку библиотеки в таком случае сравнительно низкие, а последующие процессы менее сложны из-за этапа объединения, поэтому капельные платформы обычно имеют самую высокую пропускную способность. Обычно затраты на последующее секвенирование становятся ограничивающим фактором, так что в типичных экспериментах охват довольно низок, всего несколько тысяч различных транскриптов на клетку. Одним из основных недостатков является то, что в протоколах нет возможности контролировать количество клеток, таким образом, данные подвержены смещению, что приводит к неточному отражению биологии изучаемой системы.

На сегодняшний день существует множество платформ для анализа данных секвенирования единичных клеток. Сюда включаются веб-сервисы, устанавливаемые приложения, или скрипты, работающие на установленных на компьютере средах разработки R и Python.

Для анализа данных scRNA-seq существует ряд известных программных продуктов, таких как 10x Genomics Cell Ranger [98], 10x Genomics Loupe Browser, SeqGeq (BD Biosystems) и Partek Flow (Partek). Loupe Browser открывает .cloupe-файл сформированный Cell Ranger и содержит следующую информацию:

- Уровень экспрессии генов для каждой клетки в образце.

- Различную информацию о клетках на основании экспрессии, включая проекции t-SNE и UMAP, а также дифференциальную экспрессию.

- Информацию о гене из референсного генома.

Многие эксперименты включают информацию для многих образцов обрабатываемую через GEM Chromium или через разные GEM ячейки. В таком случае используется .aggr-файл. В зависимости от дизайна эксперимента информация может поступить от того же набора клеток, клеток

разных тканей и разных временных точек, но одного организма, или клеток от разных организмов. Когда Cell Ranger count обрабатывает данные от одного образца с одной GEM ячейки. Cell Ranger multi обрабатывает данные нескольких образцов в одной GEM ячейке. Cell Ranger aggregate обрабатывает данные от многих образцов используя несколько образцов из нескольких запусков Cell Ranger count производя анализ обобщенных данных. Также aggr обобщает выводы нескольких образцов через один или несколько запусков Cell Ranger multi. Aggr формирует выходные файлы которые содержат всю информацию из каждого входного рабочего алгоритма, собранных в один выходной файл для удобного многообразцового анализа. Суффикс GEM ячейки каждого баркода обновляется для избежания совпадения баркодов.

Коммерческие программы легки в использовании и интуитивно понятны для исследователя, но в них меньше возможностей для гибкого анализа. Bioconductor – проект, на котором выложено программное обеспечение с открытым исходным кодом предоставляет несколько мощных инструментов для анализа данных высокопроизводительного секвенирования, например Scanpy [99] и Scater [100]. Одним из самых популярных пакетов для полной обработки данных секвенирования единичных клеток является Seurat [101]. Seurat - это R-пакет, предназначенный для контроля качества, анализа и исследования данных секвенирования РНК единичных клеток. Seurat дает возможность идентифицировать и интерпретировать источники неоднородности на основе транскриптомных измерений одиночных клеток и интегрировать различные типы данных одиночных клеток [102, 103, 104, 105].

Файл, создаваемый старыми версиями R — программы статистического анализа и создания диаграмм. Хранит объекты статистики (функции, значения), создаваемые пользователем с помощью подсказки R пока открыта

программа. Используется в качестве старого расширения для R, которая использует в настоящее время расширение .RDATA.

Asc-Seurat [106] представляет собой веб-приложение основанное на пакете Shiny. Приложение объединяет в себе Seurat, Dynverse и BioMart. Используя Seurat есть возможность изучать данные scRNA-seq популяций клеток для выявления биохимических паттернов, отражающих тип клеток образца, а также определить маркеры и дифференциально экспрессирующиеся гены в каждой клетке или кластере. Благодаря встроенному пакету Dynverse в Asc-Seurat есть возможность использования большого количества моделей для построения и визуализации траекторий развития клеток. С помощью третьего компонента BioMart есть возможность немедленного функционального анализа по терминам GO для нескольких организмов.

Однако, необходимость обладать навыками разработки в R или Python создают непреодолимые трудности в использовании подобных инструментов. Проект Galaxy [107] был создан с той же целью, что и Bioconductor, но разработчики облегчили использование имеющихся там инструментов и сделали их доступными для ученых, не имеющих знаний в биоинформатике.

Лабораторией Харченко П. был разработан целый ряд программных решений для анализа данных scRNA-seq. Среди них наиболее популярными стали Conos [108] (для анализа нескольких образцов), Pagoda2 [109] (полный процессинг данных scRNA-seq), dropEst [110] (демультиплицирование данных scRNA-seq).

Принципиально новые возможности для анализа данных scRNA-seq открыли методы, основанные на важном наблюдении. Изучая данные scRNA-seq, полученных с помощью протоколов SMART-seq2, STRT/C1, inDrop и 10x Chromium, было обнаружено, что 15–25% прочтений содержат несплайсированные интронные последовательности, по сравнению с предыдущими наблюдениями в секвенировании bulkRNA (14,6%) и scRNA-

seq (~20%). Velocyto [111] (расчет RNA velocity [112]), scVelo [113] (подсчет соотношения сплайсированных и несплайсированных форм). Новые методы позволили заглядывать в недалекое будущее единичных клеток, получая информацию о направлении их развития. По сути, получение данных об RNA-velocity частично решило вопрос о необходимости получения биологических или технических повторений. Эталонным способом изучения направления дифференцировки клеток считается секвенирование транскриптома клеток с разницей в часы или дни, для того чтобы можно было сравнить транскрипционные профили клеток и определить направление их дифференцировки или переключение метаболизма. Соотношение сплайсированных форм РНК к несплайсированным оценивает динамику изменения количества активных транскриптов одного гена, учитывая то, что экспрессия может усиливаться (идет активная транскрипция), снижаться (транскрипция постепенно снижается, трансляция идет более активно) или исчезать (транскрипция прекращается, трансляция снижается, усиливается деградация). Поскольку соотношение этих форм в момент времени является довольно характерным показателем, с помощью математических методов можно оценить дальнейшее изменение этих соотношений, и, соответственно, предсказать, какой транскрипционный профиль будет характерен для клетки через некоторое время.

Еще одним интересным подходом к анализу данных scRNA-seq является метод построения траекторий развития. Такой подход возник раньше RNA-velocity. Биологический смысл метода предельно прост – все клетки образца выстраиваются в некую последовательность, где каждая следующая клетка располагается после предыдущей на основании схожести паттерном экспрессии. Таким образом получают траектории, которые могут быть линейными, разветвленными и т.д. Интересно, что ни метод траекторий, ни метод RNA-velocity по отдельности не могут с достаточной достоверностью предсказать направление дифференцировки клеток. Однако совместное

использование данных методов дает результаты, стабильно подтверждаемые экспериментально.

Бурное развитие и повсеместное применение технологий искусственного интеллекта не обошло и задач по анализу данных scRNA-seq. Если рассматривать весь процессинг данных scRAN-seq, можно выделить наиболее проблемные этапы, где применение стандартных математических методов не всегда дает желаемый результат. Происходит это по причине повсеместного использования универсальных математических методов в абсолютно разных сферах и при работе с абсолютно различными по происхождения и смыслу данными. Процессы в живом организме и отдельной клетке отличаются безграничной вариативностью и применение стандартных математических подходов не всегда оправдывает свое назначение. Попытки использования нейронных сетей и методов машинного обучения для решения задач кластеризации и типирования клеток показывают результаты, заметно отличающиеся достоверностью по сравнению с классическими подходами.

scDeepCluster [114] метод кластеризации, основанный на Лувенском алгоритме, обучаемый без учителя. В таком случае нет деления на тренировочную и тестовую выборку. Автоэнкодер учится минимизировать разницу между входными и выходными векторами. То есть, стремится выдавать тот же вектор, что получил на вход. Автоэнкодер много раз проходит по всему набору данных, изменяя веса и формируя некое низкоразмерное представление данных на своем скрытом слое. Это представление затем используется для кластеризации тех же самых данных. Для кластеризации нового набора данных необходимо снова обучить на нем автоэнкодер, взять представление со скрытого слоя и запустить алгоритм кластеризации.

scCapsNet-mask [115] представляет собой обновленную версию scCapsNet, которая использует маску для облегчения задачи интерпретации

модели. Чтобы оценить эффективность маски scCapsNet, были проведены эксперименты на двух наборах данных scRNA-seq. Результаты экспериментов на двух наборах данных scRNA-seq показывают, что scCapsNet-mask может ограничивать коэффициенты связи и внутренние параметры модели. Следовательно, scCapsNet-mask сохраняет достоинства высокой точности классификации и высокой интерпретируемости оригинальной scCapsNet, а также имеет преимущества автоматической обработки и легкой интерпретации. Во-первых, scCapsNet-mask может оценить судьбу клеток с менее дифференцированными состояниями. После определения судьбы можно было установить псевдовременной порядок клеток для каждого направления развития. Следуя этому псевдовременному порядку, гены, специфичные для направления, демонстрируют характер постепенного увеличения экспрессии, а гены, связанные с гематопоэтическими стволовыми клетками, демонстрируют постепенное снижение экспрессии. Во-вторых, scCapsNet-mask можно применять для определения типа клеток в пространственной транскриптомике. Обучаясь на данных scRNA-seq, пространственная карта предсказанных типов клеток, созданная моделью scCapsNet, согласуется с картой, созданной RCTD (устойчивая декомпозиция типа клетки), и анатомической структурой гиппокампа мыши с гораздо меньшими затратами времени и вычислительных ресурсов.

1.6 Биоинформатический анализ данных scRNA-seq

1.6.1 Обработка сырых данных scRNA-seq

При объединении несколько библиотек образцов для секвенирования на одной дорожке проточной кюветы, для снижения стоимости секвенирования, образцы демультиплексируются по их индексу на этапе создания fastq-файлов из BCL. BCL (файл данных базовых вызовов) - двоичный файл с необработанными данными секвенирования, генерируемый секвенаторами. Индексы образцов - это "штрих-коды" для

мультиплексированных образцов, которые были добавляеются в структуру рида во время подготовки библиотеки. Существует три варианта генерации файлов FASTQ из файлов BCL, все они работают для библиотек 10x Genomics Chromium. Cell Ranger mkfastq [116] - 10x Genomics оберточная программа для bcl2fastq. Mkfastq удаляет из прочтений адаптерные последовательности и уникальные молекулярные идентификаторы [117]. Также существуют еще две распротсраненных программы - Illumina's bcl2fastq и Illumina's BCL Convert.

1.6.2 Форматы файлов, используемых в анализе

Формат BAM (Binary Alignment Map) представляет собой бинарный эквивалент SAM. BAM занимает меньше места и позволяет быстрее работать с информацией, чем SAM. Однако только файлы SAM доступны для чтения как текстовые файлы.

Файлы формата CRAM являются ещё более эффективными с точки зрения занимаемого дискового пространства, чем файлы BAM. В CRAM-файл хранятся отличия прочтений от референсной последовательности, поэтому для работы с ним необходимо наличие файла с референсным геномом.

Формат SAM (Sequence Alignment Map) — это текстовый формат для хранения биологических последовательностей, выровненных по эталонной последовательности, также называемой референсной. Этот формат широко используется для хранения таких данных, как фрагменты нуклеотидных последовательностей (иначе называемых чтениями, прочтениями или ридами), полученные с помощью технологии секвенирования нового поколения. Чаще всего SAM получают в результате картирования прочтений из файла FASTQ на последовательность референсного генома. Формат поддерживает короткие и длинные чтения (до 128 Mbp) и может включать одно или несколько выравниваний. Одно выравнивание состоит из нескольких строк, каждая из которых — выравнивание одного фрагмента.

SAM-файл может содержать заголовок, строки которого всегда начинаются с символа «@», за которым следует один из двухбуквенных кодов типа заголовка. В заголовке каждая строка разделена символом табуляции, и, кроме строк @CO, каждое поле данных соответствует формату тэг: значение, где тэг представляет собой двухсимвольную строку, которая определяет формат и содержимое значения.

Под заголовком находится раздел выравнивания. Он имеет 11 обязательных полей, содержащих такую информацию, как позиция и качество выравнивания, направление прочтения, указание на парное прочтение и др. Кроме того, возможно указание ряда опциональных полей в виде тэг: тип: значение.

Формат GTF (Gene Transfer Format) представляет собой текстовый формат, предназначенный для хранения аннотаций геномов, включая информацию о генах, экзонах, транскриптах и других функциональных элементах, связанных с генетическим материалом. Этот формат широко используется в области геномики и дает возможность представлять структурные данные о генах, полученные в результате секвенирования и аннотации геномов.

GTF-файлы имеют строгую структуру, в которой каждая строка, описывающая геномный элемент, состоит из нескольких полей, разделённых символом табуляции. Стандартно GTF имеет 9 полей, каждое из которых содержит специфическую информацию о генах и их структуре. Первые два поля представляют собой идентификацию хромосомы и аннотацию источника, за которыми следуют типы элементов, такие как "gene", "exon", "transcript" и другие. Далее идут позиции начала и конца элемента на хромосоме, а также направление ("+" или "-") и, в некоторых случаях, оценка или биндинг-реплики.

В GTF-файлах также содержится информация о тегах, которые могут использоваться для описания различных атрибутов генов и транскриптов. Эти атрибуты представляются в виде ключ-значение и могут включать

идентификаторы гена, идентификаторы транскриптов, названия генов, а также другую биологически значимую информацию.

Формат GTF является взаимозаменяемым с другими форматами аннотации, такими как GFF (General Feature Format), и широко используется в процессах анализа данных секвенирования, таких как выравнивание и количественная оценка экспрессии генов. Это делает его неотъемлемой частью геномных исследований, позволяя исследователям и биоинформатикам эффективно работать с аннотированными данными генов и транспозонов, что, в свою очередь, способствует лучшему пониманию генетических механизмов и функций.

Формат BCL (Base Call File) представляет собой специальный бинарный формат, используемый в высокопроизводительном секвенировании для хранения данных о чтениях нуклеотидов, полученных с помощью технологий секвенирования следующего поколения (NGS). В частности, формат BCL применяется с платформами Illumina и хранит информацию о базовых вызовах (base calls) и их качествах, полученных во время процесса секвенирования.

BCL-файлы содержат указания на последовательности нуклеотидов, которые были определены для каждой прочитанной базы в процессе секвенирования, а также данные о качестве этих вызовов, обеспечивая важнейшую информацию для последующего анализа. Каждая запись в BCL-файле представляет собой информацию для конкретного цикла секвенирования и включает в себя базовые вызовы для каждой активной считывающей ячейки на потоке и оценки качества для этих вызовов. Данные в формате BCL обычно хранятся в виде массивов, что делает их компактными и оптимизированными для высокоскоростной обработки.

Файл BCL не является текстовым, что отличает его от многих других форматов, таких как FASTQ или SAM. Вместо этого он основывается на бинарной кодировке, что позволяет эффективно хранить большие объемы данных, связанных с многоразовыми считываниями и сложными процессами

выравнивания. Каждое значение базового вызова кодируется в четырёхзначной системе для А, Т, Г и Ц — основ нуклеотидов, что существенно экономит место по сравнению с текстовыми репрезентациями.

Формат FASTA является текстовым форматом для хранения биологических последовательностей, таких как нуклеотидные или аминокислотные последовательности. Он широко используется в биоинформатике для представления последовательностей жизни, включая ДНК, РНК и белки. FASTA-файлы содержат как саму последовательность, так и информацию о ней, что делает этот формат полезным для передачи и анализа данных в области молекулярной биологии.

Структура файла FASTA достаточно проста. Каждая запись начинает с заголовка, который представляет собой строку, начинающуюся с символа «>». За этим символом следует идентификатор последовательности и опциональное описание, которое может содержать дополнительную информацию о последовательности, такую как место нахождения образца или название организма. Заголовок может занимать всю строку или быть продолжен на следующей строке, а сама последовательность располагается под заголовком и может быть разбита на несколько строк, что улучшает читаемость.

Формат FASTQ — это текстовый формат, который используется для хранения биологических последовательностей, полученных в процессе секвенирования с оценками качества каждой нуклеотидной базы. Он был разработан для упрощения хранения и обмена данными, полученными с помощью технологий секвенирования нового поколения, и стал стандартом в области биоинформатики. Формат FASTQ обычно генерируется во время секвенирования, в частности при получении чтений из-за применения с помощью следующего поколения технологий секвенирования.

Структура файла FASTQ включает четыре строки для каждой записи, представляющей фрагмент последовательности: первая строка начинается с символа «@», за которым следует уникальный идентификатор прочтения и

опциональное описание. Вторая строка содержит саму нуклеотидную последовательность, за которой следует третья строка, имеющая символ «+» и возможный идентификатор (пополняющий или повторяющий первый). Четвертая строка представляет собой строку кодов, которые отображают качество каждой нуклеотидной базы в последовательности, используя алгоритм, такой как Phred, для кодирования качества.

Формат HDF5 (Hierarchical Data Format version 5) представляет собой мощный, гибкий и эффективный формат для хранения и управления большими объемами данных. Он широко используется в научных, исследовательских и инженерных приложениях, позволяя пользователям организовывать, хранить и извлекать разнообразные массивы и структуры данных, включая многомерные массивы, таблицы и метаданные.

Одной из ключевых особенностей формата HDF5 является его иерархическая структура, которая позволяет организовывать данные в виде групп и подгрупп, наподобие файловой системы. Это обеспечивает удобный способ организации связанных данных, позволяя структурировать их в способ, соответствующий логике исследования. Каждый объект в HDF5, будь то группа, массив данных или атрибут, имеет уникальный идентификатор и может иметь связанные метаданные, которые помогают описать содержимое и контекст данных.

1.6.3 Контроль качества полученных данных

После получения результатов scRNA-seq, необходимо проверить качество секвенирования. Для этой задачи можно использовать инструмент FastQC. FastQC - это инструмент контроля качества данных секвенирования, который можно использовать как для bulk, так и для scRNA-seq данных.

1.6.4 Картирование на геном и транскриптом

После получения данных scRNA-seq следует картирование транскриптов единичных клеток на референсный геном. Референсный геном

представляет собой электронную базу данных нуклеотидных последовательностей, образец генома какого-либо вида. Обычно референсный геном собирается из секвенированных ДНК некоторого числа особей и не является точным набором генов конкретного организма. Наиболее известной и производительной платформой для анализа данных scRNA-seq на сегодняшний день является Chromium 10X Genomics. У данной платформы имеется ряд особенностей, отличающих ее от других платформ.

Картирование на геномную библиотеку в алгоритме Cell Ranger осуществляется с помощью выравнивателя STAR. На первом этапе работы программы используются .fastq-файл (библиотека последовательностей) и .gtf-файл (файл с аннотацией) для индексации. На втором этапе STAR картирует индексированные последовательности на геном генерируя выходные файлы .sam и .bam, содержащие информацию о статистике картирования, выравнивания транскриптов на границы между интронами и экзонами, некартированные риды и т.д.

У зрелых микроРНК отсутствует поли-А хвост, поэтому они не могут быть определены в процессе секвенирования. Пре-микроРНК образуются из их предшественников – при-микроРНК. При-микроРНК содержат 5'-кэп и 3' поли-А хвост, поэтому они могут быть транксрибированы в процессе получения библиотеки 10x. Гены микроРНК намеренно удалены из предпоготовленного референсного генома, поэтому, если возникает необходимость получить информацию о при-микроРНК, геном необходимо кастомизировать, добавить гены микроРНК.

1.6.5 Получение матрицы экспрессии генов

Перед подсчетом UMI алгоритм Cell Ranger исправляет ошибки секвенирования в последовательностях UMI. Прочтения, которые точно совпадают с транскриптомом, помещаются в группы с одним и тем же баркодом, UMI и аннотацией гена. Если две группы прочтений имеют одинаковый баркод и ген, но их UMI различаются на одну пару оснований,

то один из UMI попадает в группу из-за ошибки замены при секвенировании. В этом случае UMI менее поддерживаемой группы чтения исправляется на UMI с более высокой поддержкой.

Затем Cell Ranger снова группирует прочтения по баркоду, UMI (уже исправленному) и аннотации генов. Если две или более групп прочтений имеют одинаковый баркод и UMI, но разные аннотации генов, аннотация гена с наиболее поддерживаемыми прочтениями сохраняется для подсчета UMI, а другие группы прочтений отбрасываются. В случае ничьей для максимальной поддержки прочтений все группы прочтений отбрасываются, так как ген не может быть точно определен.

После этих двух этапов фильтрации каждый баркод, UMI, комбинация генов записывается как число UMI в нефильтрованной матрице ген-баркод. Количество прочтений, поддерживающих каждый подсчитанный UMI, также записывается в файле информации о молекулах – HDF5. Файл HDF5 содержит данные, соответствующие изучаемым молекулам, а также данные об используемых библиотеках и наборах функций.

Наборы данных HDF5 в файле информации о молекулах соответствуют столбцам таблицы. Каждая строка этой таблицы соответствует уникальному (UMI, баркод клетки, функция) кортежу, указывающему функцию, лучше всего поддерживаемую прочтением (включая дубликаты ПЦР), назначенным этому UMI и баркоду клетки.

Cell Ranger представляет собой алгоритм для анализа клеток, который позволяет лучше определять популяции клеток с низким содержанием РНК, особенно когда клетки с низким содержанием РНК смешиваются с популяцией клеток с высоким содержанием РНК.

Алгоритм состоит из двух основных этапов:

Используется порог, основанный на общем количестве UMI каждого баркода для идентификации клеток. На данном этапе определяются клетки с высоким содержанием РНК.

Затем алгоритм использует профиль РНК каждого оставшегося баркода, чтобы определить, является он «пустым» или частично содержащим клетку. На втором этапе определяются клетки с низким содержанием РНК, общее количество UMI которых может быть аналогично пустым GEM.

На первом этапе исходный алгоритм анализа клеток Cell Ranger используется для определения клеток с высоким содержанием РНК, используя порог, основанный на общем количестве UMI для каждого баркода.

На втором этапе выбирается набор баркодов с небольшим количеством UMI, которые, вероятно, представляют «пустые» части GEM. Создается модель профиля РНК выбранных баркодов. Эта модель, называемая фоновой, представляет собой полиномиальное распределение по генам. Она использует простое сглаживание Гуд-Тьюринга, чтобы обеспечить ненулевую оценку модели для генов, которые не наблюдались в репрезентативном пустом наборе GEM. Наконец, профиль РНК каждого баркода, не называемого клеткой на первом этапе, сравнивается с фоновой моделью. Баркоды, профиль РНК которых сильно не совпадает с фоновой моделью, добавляются к набору положительно определенных клеток. На втором этапе идентифицируются клетки, которые четко отличимы от профиля пустых GEM, даже если в них может быть гораздо более низкое содержание РНК, чем в самых крупных клетках в эксперименте.

Баркоды могут быть определены как связанные с клетками на основе их количества UMI или их профилей РНК. Поэтому некоторые области графика могут содержать баркоды, связанные как с клетками, так и с фоном. Насыщенность цвета диаграммы представляет локальную плотность баркодов, связанных с клетками.

В некоторых случаях набор баркодов, называемых клетками, может не соответствовать желаемому набору баркодов на основании визуального осмотра. Это можно исправить, либо повторно запустив подсчет, либо повторно проанализировав с параметром --force-cells, либо выбрав нужные баркоды из исходной матрицы баркодов-генов в последующем анализе.

Пользовательский выбор баркодов может быть получен, указав --barcodes to reanalyze.

1.6.6 Подготовка данных для вторичного анализа

После получения матриц экспрессии генов в Cell Ranger для дальнейшего анализа используются файлы features.tsv, barcodes.tsv и matrix.mtx.

Для исправления групповых эффектов при использовании различных наборов реактивов, Cell Ranger использует алгоритм, основанный на взаимных ближайших соседях [118], чтобы идентифицировать похожие субпопуляции клеток между образцами. Взаимный ближайший сосед определяется как пара клеток из двух разных образцов, которые содержатся в наборе ближайших соседей друг друга [119].

Соответствующие субпопуляции клеток между образцами используются для объединения нескольких образцов вместе [120]. Разница в значениях экспрессий между клетками в паре MNN обеспечивает оценку группового эффекта. Вектор коррекции для каждой клетки получается, как средневзвешенное значение оцененных групповых эффектов, где функция ядра Гаусса увеличивает веса совпадающих векторов, принадлежащих близлежащим точкам [121].

Оценка группового эффекта определяется для количественного измерения группового эффекта до и после коррекции. Для каждой клетки вычисляется, сколько из ее k ближайших соседей принадлежат одному и тому же образцу, и нормализуем его на ожидаемое количество таких же клеток образца, если нет пакетного эффекта. Оценка группового эффекта рассчитывается как среднее значение вышеуказанного показателя в случайно выбранных 10% от общего числа клеток. Если нет группового эффекта, ожидается, что ближайшие соседи каждой клетки будут равномерно распределены по всем пакетам, а оценка пакетного эффекта будет близка к 1.

Данные секвенирования единичных клеток часто собираются на основе нескольких экспериментов с различным временем сбора, расходными материалами и технологическими платформами. Эти различия могут привести к большим вариациям или так называемым «групповым эффектам» в данных и могут усложнить интересующие биологические вариации. Tran et al. сравнили 14 методов с точки зрения времени выполнения вычислений, способности обрабатывать большие наборы данных и эффективности пакетной коррекции при сохранении чистоты типа ячеек. Основываясь на их результатах, Harmony, Liger и Seurat 3 являются рекомендованными методами для пакетной интеграции. Из-за значительно более короткого времени выполнения, Harmony рекомендуется в качестве первого метода, который можно попробовать, а другие методы - как жизнеспособные альтернативы. Luecken et al. В своих данных, представляющих> 1,2 миллиона клеток, обнаружили, что выбор генов с высокой степенью вариабельности повышает эффективность методов интеграции данных, тогда как масштабирование подталкивает методы к тому, чтобы отдавать приоритет удалению образца над сохранением биологической вариации.

1.6.7 Снижение размерности

Задачей метода снижения размерности в анализе данных scRNA-seq является снижения размерности матрицы экспрессий генов с (клетки * гены) до их наиболее важных показателей (клетки * M) где M это число главных компонент, выбираемое исследователем (Рис. 3).

Чтобы уменьшить матрицу экспрессии генов до ее наиболее важных функций, Cell Ranger использует анализ основных компонент (PCA) для изменения размерности набора данных с (клетки x гены) на (клетки x M), где M - выбираемое пользователем число основных компонент (через num_principal_comps). В конвейере используется реализация алгоритма IRLBA на языке Python [122]. Конвейер повторного анализа позволяет пользователю дополнительно сокращать данные путем случайной

подвыборки клеток или выбора генов по их разбросу в наборе данных. Если данные содержат данные баркода функции, для PCA и последующего анализа используются только данные экспрессии генов.

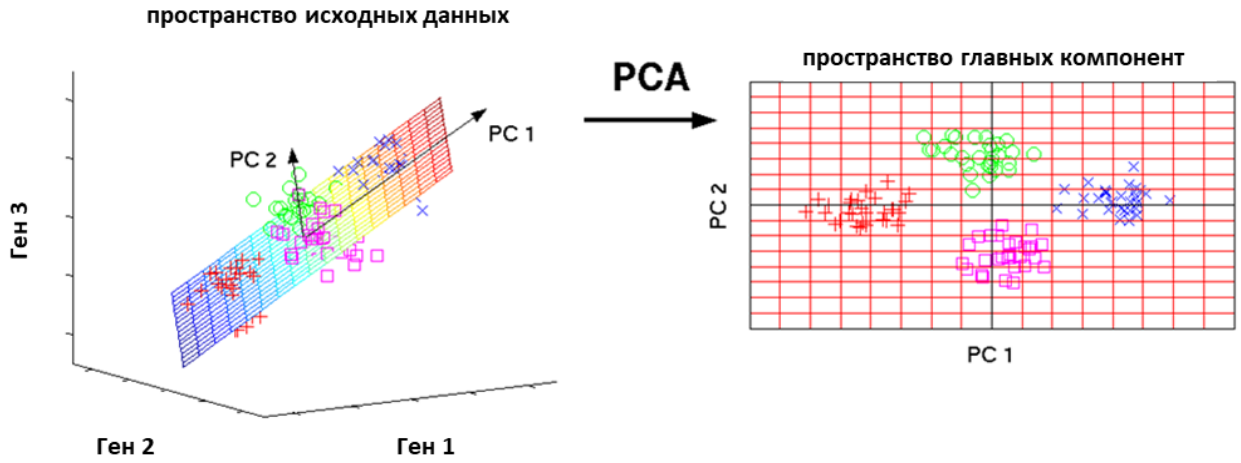


Рисунок 3. Принцип метода главных компонент [123]

Смысл уменьшения размерности в том, чтобы визуализировать многомерные данные в двумерном пространстве, в графике координат. Изначально выбираются главные компоненты, которые максимально полно описывают дисперсию имеющихся в массиве элементов. Затем для каждого элемента строится перпендикуляр к линии главной компоненты и оценивается расстояние от этого компонента до линии главной компоненты. Проекции элементов на главную компоненту затем группируются (кластеризуются) по принципу наименьшего расстояние между ними и те, расстояния между которым минимальны – группируются в один кластер, другие – в следующий кластер и так далее.

Снижение размерности - важный инструмент, необходимый для упрощения очень сложного информационного содержания в анализе данных scRNA-Seq. Правильное уменьшение размеров позволяет эффективно удалять шум и имеет решающее значение для многих последующих анализов, которые включают кластеризацию клеток или реконструкцию клонов. Sun et al. сравнили 18 различных методов уменьшения размерности с 30 общедоступными наборами данных scRNA-seq [124]. Было предположено,

что применение сложных подходов к фильтрации генов перед запуском уменьшения размерности поможет улучшить их производительность. Кроме того, обсуждается преимущество в еще более строгих подходах к фильтрации генов, поскольку они приводят к уменьшению подмножества генов и, следовательно, облегчают применение некоторых методов медленного уменьшения размерности к большим наборам данных. Основная проблема при уменьшении размерности состоит в том, чтобы сохранить глобальную структуру данных, поскольку удаление измерений аналогичным образом может скрыть некоторую информацию. Некоторые алгоритмы, такие как алгоритм scvis, пытаются преодолеть это ограничение, вычисляя низкоразмерные вложения данных scRNA-seq, сохраняя при этом глобальную структуру многомерных измерений [125]. Недавно Heiser et al. представили несмещенный фреймворк, который для сохранения определяет метрики глобальной и локальной структуры при преобразованиях уменьшения размерности [126].

Снижение размерности с точки зрения математики — это преобразование данных, состоящее в уменьшении числа переменных путём получения главных переменных. Этап снижения размерности в распространенных алгоритмах обработки данных scRNA-seq необходим для избавления от шума и автокорреляций, упрощения и ускорения последующего анализа (например, в случае кластеризацией), визуализации данных, избавления от «проклятия размерности». Анализ главных компонент (PCA), t-распределенное стохастическое встраивание соседей (t-SNE) и приближение и проекция однородного многообразия (UMAP), а также многие расширения этих трех алгоритмов обычно используются в scRNA-seq.

Методы линейного выделения признаков

Одним из самых известных методов линейного выделения признаков является PCA (Principal Component Analysis, рус. метод главных компонент). Основной идеей этого метода является поиск такой гиперплоскости, на

которую при ортогональной проекции всех признаков максимизируется дисперсия. Данное преобразование может быть произведено с помощью сингулярного разложения матриц и создает проекцию только на линейные многомерные плоскости, поэтому метод находится в категории линейных.

Задачей PCA является описание дисперсии данных. В стандартной обработке данных используют первые две компоненты, как описывающие более 98% дисперсии данных. Однако, некоторые авторы советуют использовать все возможные компоненты, принимая во внимание специфичность биологических данных. Также существует несколько иной взгляд на визуализацию данных в координатах разных компонент. Так, например, есть мнение о том, что при визуализации объекта в координатах разных компонент, есть возможность лучше рассмотреть неудачно расположенные кластеры клеток и обнаружить незамеченные [127].

Основной задачей метода главных компонент является сохранить как можно больше информации в первой компоненте, затем как можно больше во второй и так далее. Тогда отбросив некоторое количество последних компонент, содержащих мало информации, оставив главные, размерность уменьшается. Следует заметить, что после такой обработки информации отсутствует возможность смысловой интерпретации компонент.

Объясним идею метода на тривиальном примере. Пусть в исследовании в описание клеток входит только три гена. Тогда для каждой клетки имеется набор из трех чисел, соответствующих количеству гена в клетке. Эти наборы чисел можно изобразить точками трехмерного пространства. Исходная информация будет перераспределена на три компонента – по размерности задачи. Фактически это означает, что на трехмерной картинке вместо исходной прямоугольной системы координат будет выбрана другая, направив оси в направлении наибольшего разброса (дисперсии) точек.

Допустим, нужно уменьшить размерность задачи до двух. Тогда отбрасываются компоненты, содержащие наименьшее количество информации. Другими словами, исключается направление (выбор двумерной

проекции), вдоль которого разброс точек минимален, по сравнению с двумя оставшимися.

В общем случае, если необходимо получить двумерное изображение для n – мерной задачи, то сперва методами математической статистики находится число направлений, соответствующее максимальной степени разброса точек. Затем они располагаются в порядке убывания степени разброса и отбрасываются n-2 последних направления. Полученное изображение не содержит исходной информации о составе клеток, а только характеризует степень «похожести» исходных наборов генов. Разумеется, чем больше n, тем большая часть информации теряется при понижении порядка, поэтому полученное изображение возможно использовать только для формулировки гипотезы, которую следует проверить другими методами.

Однако иногда PCA может не точно отражать наши знания в области биологии по следующим причинам: a) PCA предполагает, что экспрессия генов соответствует многомерному нормальному распределению, а недавние исследования показали, что измерения экспрессии генов с помощью микрочипов следуют супергауссовскому распределению, b) PCA декомпозирует данные на основе максимизации их дисперсии. В некоторых случаях биологический вопрос может быть не связан с наибольшей дисперсией данных [128].

PCA является одним из наиболее часто используемых методов уменьшения размерности в биологических науках. Было показано, что первые несколько главных компонент тесно связаны с исходной тканью и что проецирование на первые главные компоненты дает нам информативный способ визуализации этих чрезвычайно многомерных данных. Поскольку главные компоненты упорядочены в соответствии с числом объясняемых вариаций (первая компонента объясняет наибольшую дисперсию, вторая компонента объясняет наибольшую дисперсию и т. д.), исследователи часто выбирают несколько первых главных компонент, полностью игнорируя

информацию, которая может быть скрыта в других главных компонентах [129].

Методы нелинейного выделения признаков

Одним из принципиальных отличий нелинейных методов t-distributed Stochastic Neighbor Embedding (t-SNE) и Uniform Manifold Approximation & Projection (UMAP) от PCA является то, что они стремятся сохранить сходство (и отличие) между клетками в пространстве низкой размерности в то время, как задача PCA - сохранить глобальную структуру данных. Поясним принцип их работы выбрав в качестве пространства низкой размерности плоскость, подразумевая, что для трехмерного пространства все рассуждения сохраняются без изменений.

К нелинейным методам, например, могут быть отнесены методы, отображающие исходное пространство признаков на нелинейные поверхности или топологические многообразия. Одним из таких алгоритмов является t-SNE (t-distributed Stochastic Neighbor Embedding, рус. стохастическое вложение соседей с t-распределением).

Реализация t-SNE делится на два основных этапа. Сначала t-SNE создает попарные связи между всеми точками n - мерного пространства, присваивая каждой точке с номером i число p_{ij}, равное вероятности того, что эта точка похожа на точку с номером j. Вероятность высчитывается по формуле, в которую входят значения координат точек i и j. Тогда похожие точки будут выбраны с большой вероятностью, в то время как вероятность выбора непохожих точек будет мала. В результате получается неориентированный граф с n вершинами и длинами ребер p_{ij}.

Затем t-SNE строит аналогичное распределение вероятностей q_{ij} по точкам на плоскости. Другими словами, сначала случайным образом на плоскости задается n точек (столько же точек, сколько было исследуемых клеток, но у каждой точки теперь будет не n, а только две координаты). Затем эти точки начинают двигать по плоскости, добиваясь их такого

расположения, чтобы вероятности похожести точек на плоскости q_{ij} были как можно ближе к аналогичным вероятностям $p_{ij.}$ То есть метод моделирует каждую клетку двумерной точкой таким образом, что похожие клетки моделировались близко расположенными точками, а непохожие клетки моделировались с большой вероятностью точками, далеко друг от друга отстоящими. Таким образом метод сохраняет расстояния между точками исходного n-мерного пространства. Критерием оптимальности расположения точек в двумерном пространстве является минимальное значение некоторой величины, вычисляемое после каждого изменения положения точек на плоскости и определяемой как «сумма отличий» p_{ij} от q_{ij} для всевозможных i и j.

Авторы этого метода - Лоуренс ван дер Маатен и Джеффри Хинтон [130] предлагают следующую физическую аналогию работы алгоритма: все точки соединены пружинами. Жесткость пружины, соединяющей точки i и j, зависит от разности между сходством двух точек в многомерном пространстве и двух точек на плоскости. Если систему «отпустить», через какое-то время она придет в равновесие, это и будет искомое распределение. Результирующая сила будет существенным образом стягивать точки двумерного пространства для близлежащих точек многомерного пространства, и отталкивать — для удаленных.

UMAP - был создан Лилендом Макиннесом [131] совместно с его коллегами как альтернатива t-SNE и объединяет в себе достоинства алгоритма t-SNE в плане уменьшения размерности и PCA в плане скорости. Еще одним его плюсом является то, что он пытается сохранить не только локальное, но и глобальное расстояние между точками.

По результатам исследований, у UMAP нет ограничений на размерность исходного пространства, t-SNE же подвержен так называемому «проклятию размерности», которое означает экспоненциальный рост числа вариантов при переборе положений точек на плоскости с ростом n, и, как следствие, проигрыш в производительности UMAP. Секрет UMAP в том, что в отличие

от t-SNE, который рассматривает всевозможные пары точек, UMAP сравнивает каждую точку только с k ее ближайшими соседями.

Если сравнивать между собой PCA, t-SNE и UMAP, то можно сказать, что задачи, с которыми хорошо справляется PCA, будут не менее хорошо решаться при помощи t-SNE и UMAP. Обратное в общем случае неверно.

Среди основных недостатков t-SNE указывают большие вычислительные затраты по сравнению с PCA и UMAP, а также возможное увеличение времени выполнения в случаях, когда начальные значения выбраны неудачно, так как они задаются случайным образом. Однако в этих случаях речь идет о миллионах точек. Если же их всего несколько тысяч, то это становится несущественным.

Таблица 1. Функции, используемые в пакете Seurat

Функция	**Аргументы**
Линейное снижение размерности	
RunPCA	npcs
Нелинейное снижение размерности	
RunTSNE	dims
RunUMAP	dims

В пакете Seurat есть возможность применять к данным как линейные, так и нелинейные методы снижения размерности. При указании размерности в аргументе функций RunTSNE и RunUMAP, можно дополнительно получить координаты третьей размерности и при визуализации получить более четкое понимание взаиморасположения кластеров по отношению друг к другу (Табл. 1).

Несмотря на широкое использование алгоритмов уменьшения размерности, таких как t-SNE и UMAP, у этих алгоритмов есть характеристики, которые приводят к отсутствию доверия: они не сохраняют важные аспекты многомерной структуры и чувствительны к произвольному выбору пользователя.

Другими словами, при анализе биологических данных, всегда необходимо понимать, что для их обработки и визуализации используются

математические методы, применяемые наравне с биологией и в других отраслях науки. Отсутствие понимания математического смысла каждого из этапов преобразования сырых данных вынуждает профильных специалистов относиться к получаемым результатам как к данности, не имеющей другого варианта представления. Однако, при изменении параметров, закономерности, выявленные при первом анализе, сохраняются, но с небольшими вариациями. Сохранение закономерностей является признаком наличия определенных соотношений между элементами набора данных. Навык интерпретации получаемых результатов формируется постепенно, при обсуждении выявленных закономерностей с коллегами. В большинстве случаев получаемые результаты подтверждают или отражают ожидания, но наиболее ценными являются находки, идущие вразрез с существующими представлениями, а иногда и переворачивающие их. В любом случае, каждый результат полученный биоинформатически необходимо впоследствии подтверждать экспериментально.

Вопросы и ограничения методов снижения размерности

Высокая размерность и зашумленность биологических данных

PCA предполагает, что экспрессия генов подчиняется многомерному нормальному распределению

PCA разделяет данные на основе максимизации их дисперсии

Не сохраняют важные аспекты многомерной структуры и чувствительны к произвольному выбору пользователя

Исследователи часто выбирают первые несколько главных компонент, полностью игнорируя информацию, которая может быть скрыта в других компонентах

Пользователю предоставляется возможность самостоятельно разгадать биологический смысл результатов [132]

Эти методы могут привести к тому, что пользователи будут игнорировать важную информацию, скрытую в более высоких измерениях

Из-за отсутствия интерпретируемой связи между функциями данных и низкоразмерным представлением их использование в качестве инструментов для выработки гипотез ограничено [133]

Применение только одного метода недостаточно для захвата всех важных сигналов [134]

Биологический мир кажется намного сложнее, чем мир небесной механики [135]

Снижение размерности на сегодняшний день представляет собой неотъемлемую часть обработки биологических данных. Суть данного этапа заключается в выявлении стабильных закономерностей, сопровождающихся сознательным удалением, с математической точки зрения, дублирующейся информации, в то время как в биологии необходимо учитывать и малозначительные детали. Интеграция также является дискуссионным этапом, так как до сих пор на сформулированы закономерности использовании различных методов при интеграции датасетов образцов, полученных в разное время, при различных условиях и т.д.

1.6.8 Кластеризация

Кластерный анализ предполагает выделение компактных, геометрически удаленных друг от друга групп объектов, внутри которых объекты близки (Рис. 4).

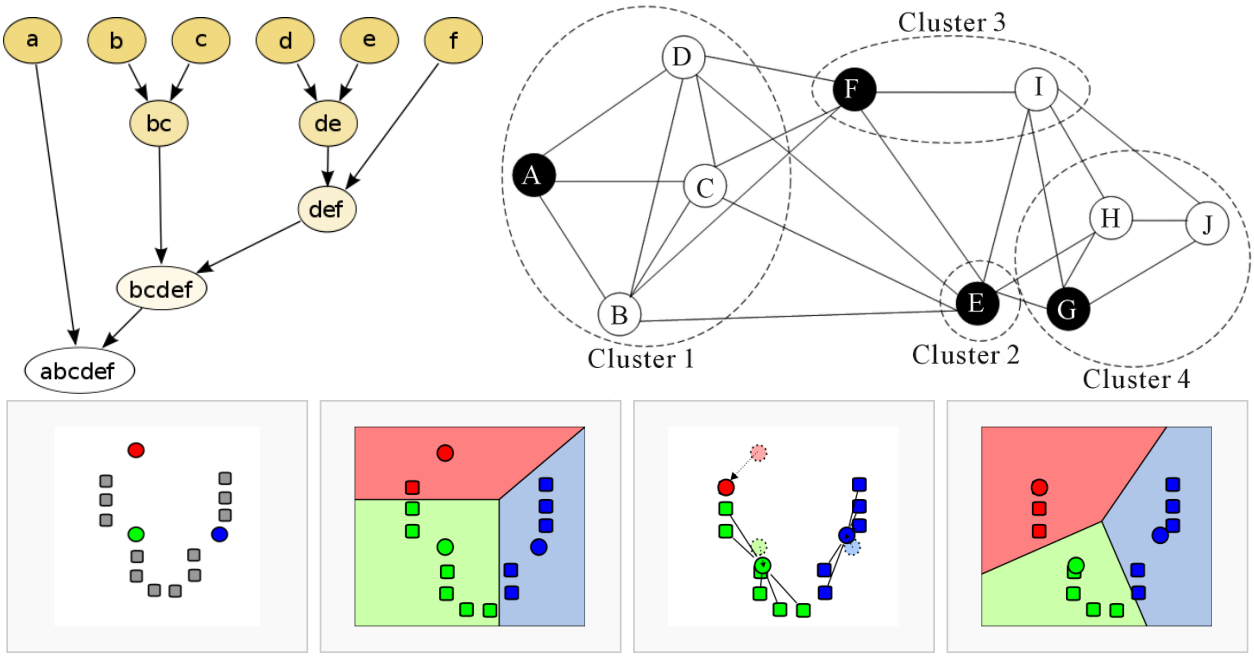


Рисунок 4. Иерархический, графовый и статистический методы кластеризации [136, 137, 138]

Кластеризация начинается с выборки объектов для кластеризации и определения множества переменных, по которым будут оцениваться объекты в выборке. Иногда из-за неоднородности единиц измерения признаков становится невозможно корректно рассчитать расстояния между точками, и тогда приходится прибегать к нормализации значений переменных. Однако следует помнить, что нормировка сильно искажает геометрию исходного пространства, что может изменить результаты кластеризации.

Кластерный анализ строится на предположении, что, имея некоторую метрику, описывающую меру сродства двух структур, и применяя специальные алгоритмы, можно разделить множество объектов на набор подмножеств [139] объектов, схожих с друг с другом в терминах выбранной метрики. Поэтому выбор метрики и алгоритма имеет решающее значение для получения корректных результатов.

Выбор расстояния между объектами является ключевым моментом исследования, от него во многом зависит окончательный вариант разбиения объектов на классы при данном алгоритме кластеризации. Однако, данный этап представляет основную сложность анализа, так как выбор метрики неоднозначен.

Евклидова метрика очень популярна и наиболее употребительна, так как отвечает интуитивным представлениям о близости объектов в пространстве. Примеры некоторых наиболее употребительных функций расстояния приведены в таблице 2 [140].

Таблица 2. Некоторые функции расстояния

Название	Формула
Линейное расстояние	$d(I_i, I_j) = \sum_{k=1}^{m} \left| I_{ki} - I_{kj} \right|$
Евклидово расстояние	$d(I_i, I_j) = \left[\sum_{k=1}^{m} (I_{ki} - I_{kj})^2 \right]^{\frac{1}{2}}$
Квадрат евклидова расстояния	$d(I_i, I_j) = \sum_{k=1}^{m} (I_{ki} - I_{kj})^2$
Обобщенное степенное расстояние Минковского	$d(I_i, I_j) = \left[\sum_{k=1}^{m} (I_{ki} - I_{kj})^p \right]^{\frac{1}{p}}$

В настоящее время известно более сотни разных алгоритмов кластеризации. Их разнообразие объясняется не только разными вычислительными методами, но и различными концепциями, лежащими в основе модели кластеризации. Наиболее популярными являются:

- модели на основе иерархии (иерархическая кластеризация) включают в себя два подхода: агломеративный и дивизимный, в зависимости от того объединяются или разделяются кластеры по мере продвижения вверх по иерархии;

- центроидные модели (k-means, k-medoids, k-medians, c-means и др. алгоритмы) представляют каждый кластер с помощью вектора средних значений. Одним из самых больших недостатков подобных алгоритмов является то, что число кластеров k необходимо знать заранее. Выбор числа кластеров может базироваться на результатах предшествующих исследований или теоретических соображениях;

- модели на основе статистического распределения, например, многомерного нормального распределения, используемого в EM-алгоритме [141];

- в моделях на основе плотности кластеры определяются как области с более высокой плотностью, по сравнению с остальной частью набора данных. Наиболее популярными алгоритмами кластеризации в данной области являются DBSCAN и OPTICS. Недостаток таких алгоритмов в том, что необходимо указывать минимальное значение плотности для определения границ кластеров [142].

Не существует объективно «правильного» алгоритма кластеризации, для конкретной задачи наиболее подходящий алгоритм кластеризации нужно выбирать экспериментально [143].

Cell Ranger использует два разных метода кластеризации клеток по сходству экспрессии, оба из которых работают в пространстве PCA.

Алгоритм кластеризации на основе графов состоит из построения разреженного графа ближайших соседей (где клетки связаны, если они находятся среди k ближайших евклидовых соседей друг к другу), за которым следует оптимизация модульности Лувена [144], алгоритм, который пытается найти сильно связанные «модули» на графе. Значение k, количество ближайших соседей, устанавливается в логарифмическом масштабе с количеством клеток. Выполняется дополнительный этап слияния кластеров: иерархическая кластеризация кластеров-медоидов в пространстве PCA и объединение пар одноуровневых кластеров, если между ними нет генов, дифференциально экспрессируемых между ними (со скорректированным методом Бенджамина-Хохберга значением p-value ниже 0,05). Иерархическая кластеризация и слияние повторяются до тех пор, пока не останется больше пар кластеров для слияния. Оптимизация модульности Лувена для кластеризации клеток ранее была использована в аналогичном пакете R Seurat.

Cell Ranger также выполняет традиционную кластеризацию k-средних по диапазону значений k, где k - заданное количество кластеров. Среди многообразия различных индексов, использующихся для валидации кластеров, наиболее популярным является индекс - Дэвиса Болдина, который используется в программном конвейнере Cell Ranger.

Метод динамических ядер (k-means) – наиболее простой и широко используемый итеративный алгоритм кластеризации без учителя, разделяющий множество данных на k кластеров, расположенных на возможно больших расстояниях друг от друга. Данный алгоритм состоит из следующих этапов:

1. Случайный выбор k точек, являющихся начальными «центрами масс» кластеров, которые могут представлять собой либо k из I объектов, либо случайные точки. Под «центром массы» кластера понимается точка в пространстве характеристических векторов со средними для данного кластера значениями характеристики;

2. Каждый объект приписывается к кластеру с ближайшим «центром масс»;

3. Перерасчет «центров масс» согласно текущему набору объектов, входящих в состав кластеров;

4. Если критерий остановки алгоритма не удовлетворен, алгоритм возвращается ко второму шагу [145, 146].

Задача кластеризации в биологии также является очень ответственным этапом. Кластеризация от классификации отличается тем, что при кластеризации исследуемые элементы группируются по набору признаков. При анализе биологических клеточных образцов элементами являются клетки, а признаками – гены.

Для начинающих исследователей получение разного количества кластеров при анализе данных может стать довольно трудной задачей. Существует мнение о том, что чем большее количество кластеров удастся получить, тем больше информации о наличии в образце различных типов

клеток или клеток, находящихся на разных этапах дифференцировки или клеток, в которых активированы на момент секвенирования специфические биохимические процессы. При освоении метода действительно может так показаться.

Для того, чтобы ответить на вопрос имеет ли смысл увеличение количества кластеров необходимо провести первое наблюдение. На рисунке 5 с увеличением разрешения увеличивается количество кластеров. В данном случае оно увеличивается с 2 до 9. Однако уже даже при анализе количества кластеров заметно, что на рисунке b, в 0 кластере появляется 1 и 2 кластеры, на рисунке c к ним добавляется 3 кластер, на рисунке d – в центре остатка 0 кластера выделяются еще 1 и 3 кластеры и т.д. В процессе увеличения значения разрешения какие-то кластеры остаются неразделенными в течение нескольких шагов, а какие-то начинают делиться уже на втором шаге. Такое наблюдение является биологически важным, так как дает исследователю выявить стойкие к кластеризации кластеры, которые, скорее всего содержат специфические признаки и их можно идентифицировать как клетки особого типа, этапа дифференцировки или с характерными биохимическими процессами.

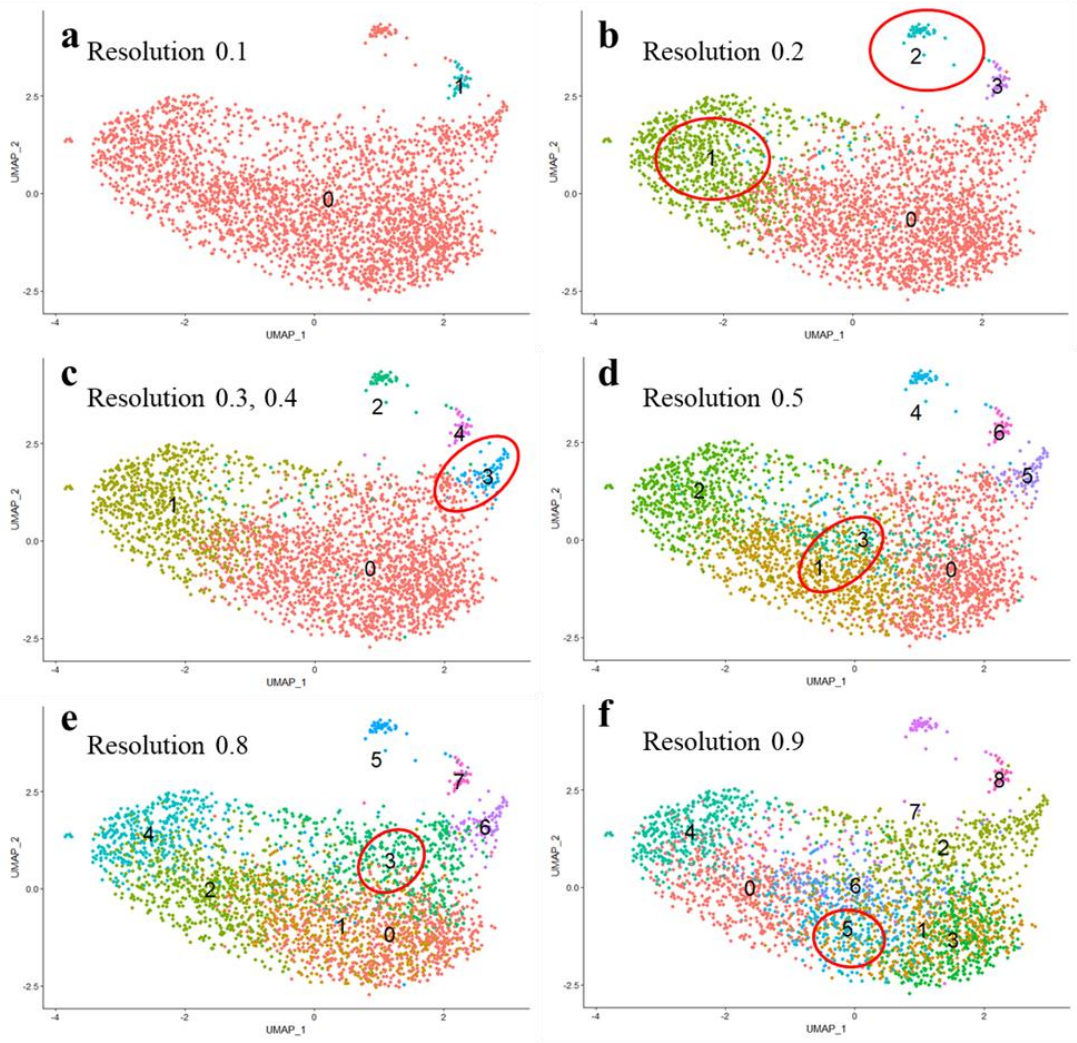


Рисунок 5. Изменение количества кластеров в зависимости от разрешения

Вторым наблюдением является изменение количества клеток в кластерах (Табл. 3). При количестве кластеров два – все клетки распределяются на две группы - 2893 и 34 клетки. При 4 кластерах – 0 кластер, содержащий 2893 клетки разделяется на три с количеством клеток в них – 200, 806 и 77. При этом в четвертом кластере остается 35 клеток, что на одну клетку меньше, чем при разделении клеток образца на два кластера. Таким образом, можно сказать, что биологически значимые клетки располагаются в кластерах, которые не изменяют своего размера при увеличении показателя разрешения.

Таблица 3. Количество клеток в кластерах при различных значения показателя разрешения

Разрешен	Количеств	Число клеток в кластере

ие	о кластеров	0	1	2	3	4	5	6	7	8
0.1	2	2893	34							
0.2	4	2009	806	77	35					
0.3	5	1935	805	79	73	35				
0.4										
0.5	7	1066	875	589	207	83	72	35		
0.6										
0.7										
0.8	8	686	598	561	500	384	82	81	35	
0.9	9	540	536	477	401	386	272	193	87	35
1										

Resolution 0.1								
DCN	CD36							
PTX3	FABP4							
SFRP2	CRYAB							
IGF1	POSTN							
MMP2	TIMP3							
LUM	IGFBP5							
COMP	COL4A1							
PDGFRA	COL4A2							
PLPP3	NDUFA4L2							
EFEMP1	MYLK							
Resolution 0.2								
LUM	ACTA2	LRRC75A	CD36					
AKR1C1	TAGLN	SCD	FABP4					
IGF1	MYL9	TGFBI	POSTN					
PLPP3	TPM1	COL4A1	CRYAB					
TWIST1	CALD1	IGFBP5	TIMP3					
RAB31	PALLD	FN1	IGFBP5					
GPX3	RGCC	LOX	COL4A1					
SAT1	MYLK	POSTN	COL4A2					
GJA1	MFAP5	FADS1	NDUFA4L2					
SRPX	ACTB	THBS1	MYLK					
Resolution 0.3-0.4								
LUM	ACTA2	LRRC75A	FABP4	CD36				
AKR1C1	TAGLN	TGFBI	FABP5	FABP4				
IGF1	MYL9	COL4A1	G0S2	POSTN				
PLPP3	TPM1	SCD	LPL	CRYAB				
TWIST1	CALD1	IGFBP5	IGFBP5	TIMP3				
MMP3	PALLD	FN1	APOE	IGFBP5				
RAB31	MYLK	LOX	PNPLA2	COL4A1				
SAT1	RGCC	FADS1	CEBPA	COL4A2				
GJA1	MFAP5	POSTN	ACACB	NDUFA4L2				
DCN	ACTB	THBS1	CRYAB	MYLK				
Resolution 0.5-0.7								
IGF1	DKK1	ACTA2	CHRDL1	LRRC75A	FABP4	CD36		
MMP3	PTX3	TAGLN	PLPP3	COL4A1	FABP5	FABP4		
PRRX1	IL1RL1	MYL9	GPX3	SCD	G0S2	POSTN		
TWIST1	MT1E	TPM1	LUM	TGFBI	LPL	CRYAB		
COL6A3	MT1X	CALD1	CFD	IGFBP5	IGFBP5	TIMP3		
OLFML2B	MT1M	MYLK	AKR1C1	FN1	APOE	IGFBP5		
SELENOP	TSC22D3	PALLD	APCDD1	LOX	PNPLA2	COL4A1		
ZFP36L2	COMP	ACTB	DCN	FADS1	CEBPA	COL4A2		
HMCN1	MORF4L2	CAV1	SPON2	COL1A1	ACACB	NDUFA4L2		
SFRP2	ITGBL1	TPM2	MT2A	THBS1	CRYAB	MYLK		
Resolution 0.8								
MMP3	AKR1C1	PTX3	PLIN2	ACTA2	LRRC75A	FABP4	CD36	
SFRP2	PLPP3	DKK1	C7	TAGLN	SCD	FABP5	FABP4	
COL6A3	CHRDL1	IL1RL1	APOE	MYL9	TGFBI	G0S2	POSTN	
PRRX1	LUM	RGCC	CLDN11	TPM1	FN1	LPL	CRYAB	
HMCN1	IGF1	MFAP5	RSPO3	MYLK	COL4A1	IGFBP5	TIMP3	
AL139393.2	MTR	COL1A1	MMP14	CALD1	FADS1	APOE	IGFBP5	
ASS1	GPX3	COMP	SCD	PALLD	LOX	PNPLA2	COL4A1	
SCRG1	DCN	MT1E	ADAM12	CAV1	IGFBP5	CEBPA	COL4A2	
COL3A1	RAB31	ELN	ADH1B	ACTB	COL1A1	ACACB	NDUFA4L2	
CXCL3	MGP	SCG2	CRYAB	CNN1	THBS1	CRYAB	MYLK	
Resolution 0.9								
PTX3	COL6A3	FABP4	IGF1	ACTA2	DKK1	CHRDL1	LRRC75A	CD36
DKK1	SFRP2	FABP5	MMP3	TAGLN	IL1RL1	PLPP3	SCD	FABP4
RGCC	ASS1	G0S2	GALNT15	MYL9	MORF4L2	GPX3	COL4A1	POSTN
IL1RL1	SCRG1	IGFBP5	LUM	TPM1	AKR1C1	CFD	TGFBI	CRYAB
COL1A1	MT1E	APOE	TWIST1	MYLK	MTR	LUM	FADS1	TIMP3
MFAP5	PRRX1	LPL	MGP	CALD1	RASSF4	AKR1C1	FN1	IGFBP5
MT1E	COL3A1	CRYAB	RAB31	PALLD	RAB31	SPON2	LOX	COL4A1
COMP	MT1X	SCD	TNFSF10	CAV1	SULF2	APCDD1	IGFBP5	COL4A2
SCG2	HMCN1	C7	FBXO32	ACTB	TSC22D3	SAA1	COL1A1	NDUFA4L2
ELN	MT2A	PLIN2	GJA1	CNN1	BEX3	DCN	THBS1	MYLK

Рисунок 6. Списки первых 10 генов каждого кластера при различных разрешениях

Третьим наблюдением являются списки наиболее представленных генов в каждом кластере при различных разрешениях (Рис. 6). Важным является

ответ на вопрос – имеют ли вновь появляющиеся при увеличении разрешения кластеры биологическое значение, или это увеличение связано с указанием желаемого количества кластеров. На рис. приведены списки первых 10 наиболее представленных генов в каждом кластере в различных разрешениях. Кластеры с похожими (но не аналогичными) наборами генов выделены одинаковыми цветами. При визуальном анализе данной таблицы видно, что некоторые кластеры наследуются с самого начала кластеризации, какие-то появляются на определенном разрешении. В какой-то момент наблюдается эффект искусственного разделения биологически значимого кластера на два менее значимых. Под значимостью в данном случае необходимо понимать отношение всех генов кластера к определенному процессу, а уменьшение значимости – появление этих генов в двух соседних кластерах.

Для кластеризации в пакете Seurat применяются предложенные в 2015 году графовые алгоритмы [147, 148].

В пакете Seurat для поиска ближайших соседей используется функция **FindNeighbors()**, в основе которой лежит построение KNN-графа, основанного на евклидовом расстоянии в размерности главных компонент и улучшение весов ребер между двумя любыми двумя клетками, основанное на перекрывании в их локальных окрестностях (коэффициент Жаккара). Для поиска ближайших соседей можно использовать один из двух представленных методов - “rann” [149] и “annoy”. Rann находит k ближайших соседей для каждой точки в данном наборе данных за время O (N log N) с использованием библиотеки ANN Arya и Mount. Пакет Annoy содержит 4 метрики. **Евклидова метрика** (L2) (евклидово расстояние) — расстояние между двумя точками евклидова пространства, вычисляемое по теореме Пифагора. В анализе данных **косинусное сходство** — это мера сходства между двумя ненулевыми векторами, определенными в пространстве внутреннего произведения. Косинусное сходство — это косинус угла между векторами; то есть это скалярное произведение векторов, деленное на

произведение их длин. Отсюда следует, что подобие косинусов зависит не от модулей векторов, а только от их угла. **Расстояние городских кварталов** (L1) — метрика, введенная Германом Минковским. Согласно этой метрике, расстояние между двумя точками равно сумме модулей разностей их координат. Название «манхэттенское расстояние» связано с уличной планировкой Манхэттена. **Расстояние Хэмминга** (кодовое расстояние) — число позиций, в которых соответствующие символы двух слов одинаковой длины различны.

Для кластеризации применяется оптимизация модулярности или алгоритм Лувейна или SLM [150] для итеративного объединения клеток в группы с целью оптимизации стандартной функции модулярности. Эту операцию выполняет функция **FindClusters()**, где можно изменять аргумент resolution или разрешение (зернистость) (Табл. 4) Сначала идентифицируются кластеры клеток с помощью алгоритма кластеризации на основе модульной оптимизации общего ближайшего соседа (SNN). После этого вычисляется k-ближайших соседей и строится граф SNN. Затем оптимизируется функция модульности для определения кластеров [151]. Алгоритм оптимизации модульности (1 = исходный алгоритм Лувена; 2 = алгоритм Лувена с многоуровневым уточнением; 3 = алгоритм SLM; 4 = алгоритм Лейдена). Лейдену требуется leidenalg python [152].

Таблица 4. Функции используемые в Seurat

Функция	**Аргументы**
FindNeighbors	dims
	k.param
	nn.method = “rann”, “annoy”
	annoy.metric = “euclidean”, “cosine”, “manhattan”, “hamming”
FindClusters	resolution
	Original Louvain algorithm
	Louvain algorithm with multilevel refinement
	SLM algorithm
	Leiden algorithm

В результате кластеризации в отличие от классификации анализируемые признаки элементов объединяются в группы согласно схожести нескольких признаков. Применительно к биологии это имеет очень большое значение, так как правильно подобранные параметры используемых функций и биоцентрическое восприятие получаемых результатов позволяют выявить неизвестные до этого момента закономерности. В частности, в биологии, кластеру соответствует субпопуляции клеток, схожих по биологическому процессу, стабильно экспрессирующему мембранному маркеру или количеству прочтенных транскриптов. Поэтому, помимо автоматического или ручного типирования полученных кластеров рекомендуется получать серию проекций с различным количеством кластеров, оценивать количество клеток в них и разницу в дифференциально экспрессирующихся генах. Это поможет исключить неверные гипотезы, появляющиеся в процессе интерпретации.

При использовании методов в Seurat можно добиться улучшения визуализации, изменения количества кластеров. Однако увеличения количества кластеров вовсе не означает выявление новых подтипов клеток. Алгоритм искусственно разделяет уже существующие кластеры на более мелкие при увеличении разрешения. Понятно, что это имеет только математическое значение. Если же подходить к интерпретации полученных результатов с точки зрения биологии, то это не имеет никакого смысла, так как дифференциально экспрессирующиеся гены, определяющие отличия одного кластера от другого, не имеют никакого биологического смысла.

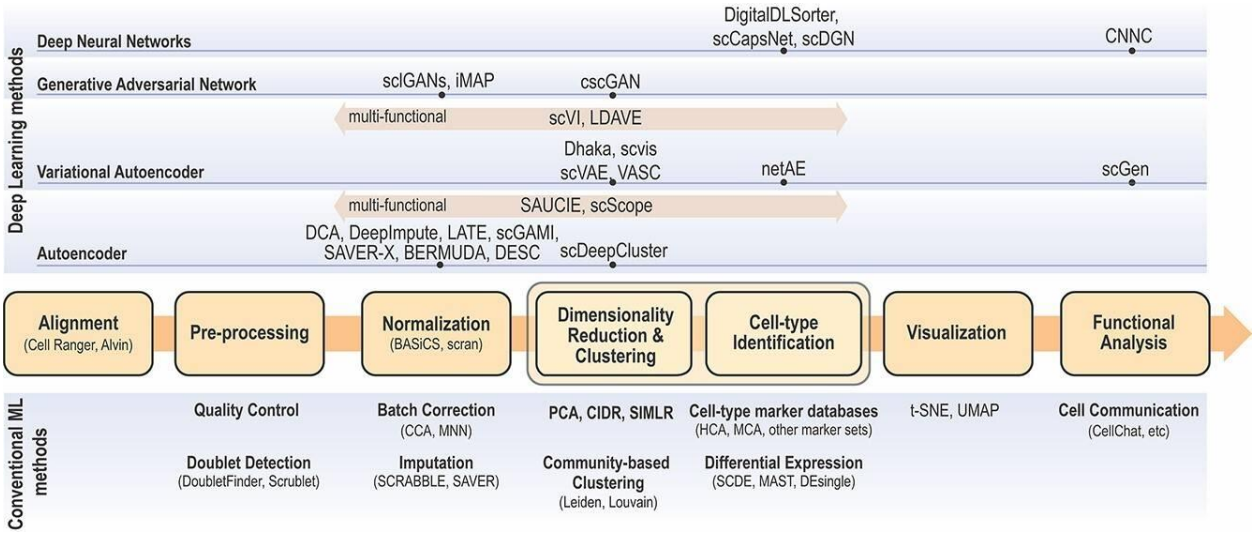


Рисунок 7. Существующие инструменты на основе методов машинного обучения и нейронных сетей для анализа данных scRNA-seq [153]
Conventional ML methods – обычные методы машинного обучения; **Deep Learning methods** – методы глубокого машинного обучения; **Deep Neural Networks** - нейронная сеть глубокого обучения; **Generative Adversarial Network** - Генеративно-состязательная нейросеть; **Variational Autoencoder** - Вариационный автоэнкодер; **Autoencoder** – Автоэнкодер; **Alignment** – картирование; **pre-processing** – предварительная обработка данных; **Normalization** – нормализация; **Dimensionality Reduction & Clustering** – снижение размерности и кластеризация; **Cell-type identification** – определение типа клетки; **Visualization** – визуализация; **Functional Analysis** – функциональный анализ

В данном случае получить принципиально другие результаты позволяет совершенно другой подход – использование для этих целей нейронных сетей (Рис. 7). Благодаря особенностям архитектуры и работы нейросетей, можно выявить гораздо более мелкие различия между клетками.

В качестве примера использования такого подхода рассмотрим scDeepCluster (Рис. 8) [154]. В статье про scDeepCluster авторы предлагают использовать автоэнкодер. Это тип архитектуры нейронной сети, который позволяет получить "сжатое" представление данных в пространстве низкой размерности, а также позволяет удалять шумы. Эти представления (вектора) далее подаются на вход алгоритму кластеризации. Данные для обучения не потребуются, так как обучение проходит без учителя. В таком случае нет деления на тренировочную и тестовую выборку. Автоэнкодер учится

минимизировать разницу между входными и выходными векторами. То есть, в идеале он стремится выдавать тот же вектор, что получил на вход. Автоэнкодер много раз проходит по всему набору данных, изменяя веса и формируя некое низкоразмерное представление данных на своем скрытом слое. Это представление и используется для кластеризации тех же самых данных. Для кластеризации нового набора данных нам нужно будет снова обучить на нем автоэнкодер, взять представление со скрытого слоя и запустить алгоритм кластеризации (в scDeepCluster это Лувенский алгоритм).

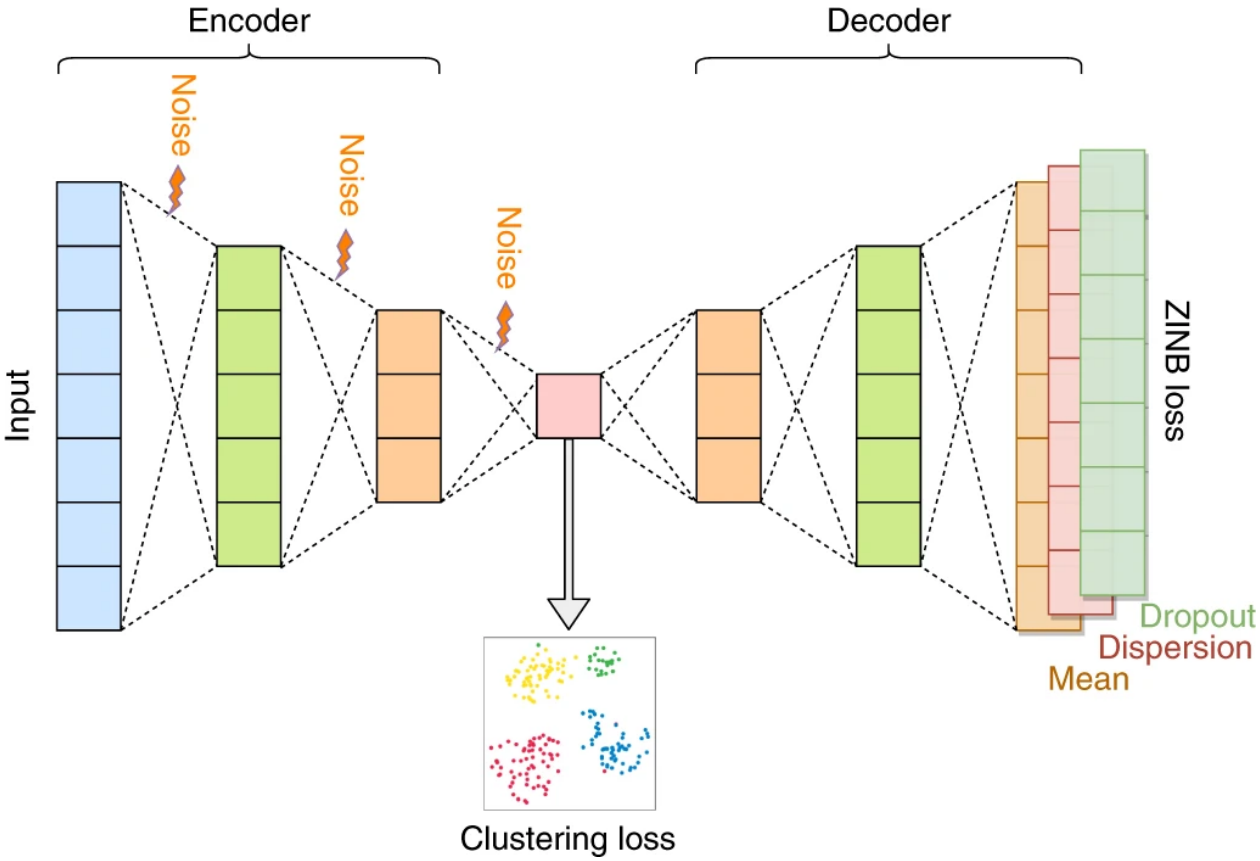


Рисунок 8. Архитектура сети scDeepCluster [155]
Input – входные данные нейронной сети; **Encoder** – нейронное кодирование; **Decoder** – нейронное декодирование; **Noise** – Шум; **ZINB loss** - отрицательное биномиального распределения с избытком нулей; **Mean** – основной выходной слой; **Dispersion** – выходной слой со значениями дисперсии; **Dropout** – исключение (метод регуляризации искусственных нейронных сетей, предназначен для уменьшения переобучения сети за счет предотвращения сложных коадаптаций отдельных нейронов на тренировочных данных во время обучения); **Clustering loss** – функция потерь при кластеризации

Существующие вопросы и ограничения методов кластеризации приведены ниже.

Вопросы и ограничения методов кластеризации

Многие алгоритмы кластеризации являются универсальными в том смысле, что их можно применять к любым типам данных, содержащими меру расстояния между точками данных. Из-за большого количества генов, анализируемых в scRNA-seq, то есть высокой размерности, расстояния между точками данных (то есть клетками) становятся одинаковыми, что известно как «проклятие размерности» [156].

Поиск универсального метода кластеризации, подходящего для всех ситуаций, может быть бесполезным, поскольку было показано, что один алгоритм не может достичь всего диапазона желаемых свойств [157]. Одним из недостатков большинства методов кластеризации является то, что они будут разделять данные независимо от того, присутствуют ли в них какие-либо биологически значимые группы [158].

1.6.9 Дифференциальная экспрессия

Для идентификации генов, экспрессия которых специфична для каждого кластера, Cell Ranger проверяет для каждого гена и каждого кластера, отличается ли среднее значение внутри кластера от среднего значения вне кластера.

Чтобы найти дифференциально экспрессирующиеся гены между группами клеток, Cell Ranger использует быстрый и простой метод sSeq [159], который использует отрицательный биномиальный точный тест. Когда количество становится большим, Cell Ranger переключается на быстрый асимптотический бета-тест, используемый в edgeR [160]. Для каждого кластера алгоритм запускается в этом кластере по сравнению со всеми другими клетками, генерируя список генов, которые дифференциально экспрессируются в этом кластере относительно остальной части образца.

Реализация Cell Ranger немного отличается. В статье sSeq авторы рекомендуют использовать определение размера библиотеки DESeq на основе среднего геометрического. Вместо этого Cell Ranger вычисляет относительный размер библиотеки как общее количество UMI для каждой клетки, деленное на среднее количество UMI для каждой клетки. Как и в случае с sSeq, нормализация подразумевается в том смысле, что параметр размера библиотеки для каждой клетки включен как фактор в вычисления вероятности точного теста.

1.6.10 Интеграция scRNA-seq датасетов

Совместный анализ двух или более наборов данных scRNA-seq создает особые сложности. В частности, определение популяций клеток, присутствующих в нескольких наборах данных, может быть проблематичным при стандартных рабочих процессах. Seurat v4 включает набор методов для сопоставления (или «выравнивания») совокупностей общих клеток в наборах данных. Эти методы сначала идентифицируют пары клеток из перекрестных наборов данных, которые находятся в согласованном биологическом состоянии («якоря»), могут использоваться как для коррекции технических различий между наборами данных (т.е. коррекция эффекта партии), так и для выполнения сравнительного анализа scRNA-seq в экспериментальных условиях.

Цели интеграции:

- Создание «интегрированного» массива данных для последующего анализа

- Определение типов клеток, которые присутствуют в обоих наборах данных

- Получение маркеров клеточных типов, консервативных как для контрольных, так и стимулированных клеток.

- Сравнение наборов данных для изучения специфичного для определенных типов клеток ответов на стимуляцию.

Определяя общие источники вариаций между наборами данных, CCA (канонический корреляционный анализ) хорошо подходит для определения якорей, когда типы клеток сохраняются, но существуют очень существенные различия в экспрессии генов в разных экспериментах. Таким образом, интеграция на основе CCA обеспечивает интегративный анализ, когда экспериментальные условия или сопутствующие заболевания вызывают очень сильные сдвиги экспрессии при интеграции наборов данных по модальностям и видам. Однако интеграция на основе CCA также может привести к чрезмерной коррекции, особенно когда большая часть клеток не перекрывается в наборах данных.

Вместо использования канонического корреляционного анализа (CCA) для определения якорей используется взаимный PCA (RPCA). При определении привязок между любыми двумя наборами данных с помощью RPCA проецируется каждый набор данных в пространство других PCA и ограничивает привязки тем же требованием взаимного соседства. Команды для обоих рабочих процессов во многом идентичны, но эти два метода могут применяться в разном контексте.

Интеграция на основе RPCA выполняется значительно быстрее, а также представляет собой более консервативный подход, при котором клетки в разных биологических состояниях с меньшей вероятностью «выравниваются» после интеграции. Поэтому RPCA используется во время интегративного анализа, когда: значительная часть клеток в одном наборе данных не имеет соответствующего типа в другом; наборы данных происходят с одной и той же платформы; существует большое количество наборов данных или клеток для интеграции.

Использование SCTransform в Seurat. Биологическая гетерогенность в данных scRNA-seq часто смешивается с техническими факторами, включая глубину секвенирования. Количество молекул, обнаруженных в каждой клетке, может значительно различаться между клетками, даже в пределах одного и того же типа клеток. Интерпретация данных scRNA-seq требует

эффективной предварительной обработки и нормализации для устранения этой технической изменчивости. В 2019 была представлена основа моделирования для нормализации и стабилизации дисперсии данных молекулярного подсчета из эксперимента scRNA-seq [161]. Эта процедура устраняет необходимость в эвристических шагах, включая добавление псевдопрочтений или логарифмическое преобразование, и улучшает общие последующие аналитические задачи, такие как определение вариабельных генов, уменьшение размерности и дифференциальную экспрессию.

Целями интеграции является создание интегрированного датасета для дальнейшего анализа, определение типов клеток, представленных в обоих образцах, получение маркеров типов клеток, присутствующих в обоих образцах и сравнение датасетов друг с другом с целью выявления специфических ответов на внешние воздействия (Рис. 9).

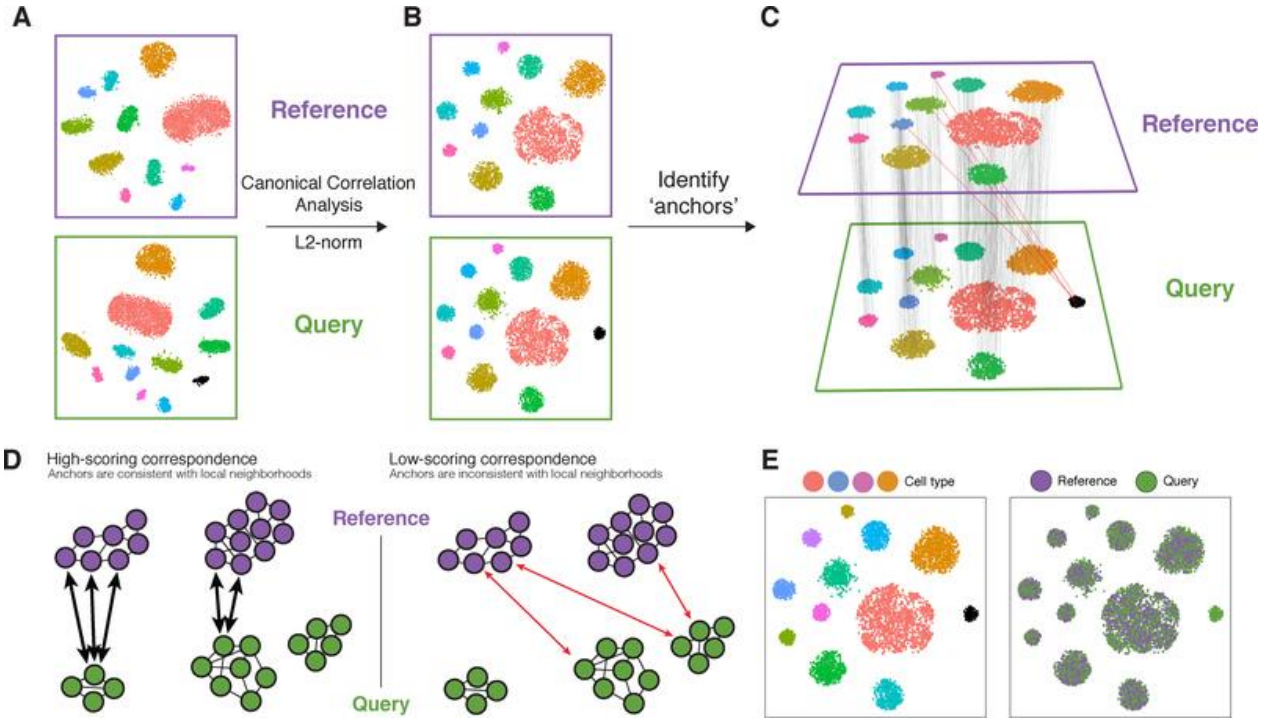


Рисунок 9. Интеграция образцов [162]

Reference – референсный образец; **Query** – анализируемые образец; **Canonical Corellation Analysis** (CCA) - канонический корреляционный анализ; **L2-norm** - Евклидова метрика (евклидово расстояние) — метрика в евклидовом пространстве — расстояние между двумя точками евклидова

пространства, вычисляемое по теореме Пифагора; **Anchors** – якоря – пары схожих друг с другом по паттерну экспрессии клетки; **High/Low-scoring correspondence** – высоко/низко-рейтинговое соответствие; **Cell type** – тип клетки

Пусть имеется два набора данных, каждый из которых представляет собой набор векторов, представляющих собой паттерны экспрессии клеток двух образцов. Не умаляя общности под интеграцией, будем понимать определение в контрольном и обработанном образцах групп клеток, паттерны экспрессии которых настолько похожи, что их можно считать клетками одного вида.

Непосредственно интеграции данных предшествует несколько этапов обработки и анализа данных. Во-первых, вне зависимости от того, будет ли выполняться интеграция, производят нормализацию данных для того, чтобы привести данные с разными диапазонами значений к единому виду, который позволит сравнивать их между собой. По-сути всё сводится к тому, что исходный набор значений сперва смещается, а потом масштабируется. После этого значения экспрессии всех генов лежат в одном диапазоне. Предварительно из наборов данных удаляются выбросы – клетки, которые, с некоторой долей вероятности являются «мусорными», и поэтому исключаются из анализа. После хорошей нормализации данных все гены должны быть равны по своему возможному влиянию, то есть никакому из них заранее нельзя отдать предпочтение или определить, что он более значим, чем остальные.

Затем определяются гены, по которым судят о степени «похожести» клеток. Если их количество не очень велико, то можно взять все. В Seurat по умолчанию это значение равно 2000 (функции FindVariableFeatures и SelectIntegrationFeatures).

Интеграция выполняется в два этапа: определение «anchors» (FindIntegrationAnchors) и интеграция датасетов (IntegrateData) (Табл. 5). «Anchors» — это пары похожих клеток в разных образцах. Сначала проводят

понижение размерности (задается параметром dims). В Seurat это можно сделать либо при помощи canonical correlation analysis (CCA), либо reciprocal PCA, либо reciprocal LSI на выбор [163]. Затем каждая клетка одного образца сравнивается с каждой клеткой другого образца. Если оказывается, что клетки похожи, то эта пара помечается как «anchor».

Поскольку уменьшение размерности происходит за счет точности, то могут появиться ложные «якоря». Поэтому после того, как было проведено сравнение всех клеток из обоих образцов, выполняется фильтрация «якорей», используя исходные непреобразованные данные, и ложные якоря удаляются. Поясним механизм фильтрации якорей. Пусть некоторый якорь «соединяет» клетки Q и R двух исходных образцов Query и Reference соответственно. В образце Reference определяют k.filter ближайших соседей (наиболее похожих клеток) клетки Q. Если среди них нет клетки R, то якорь удаляют.

Каждому оставшемуся якорю присваивается некоторый весовой коэффициент, который характеризует «качество» данного якоря. Поясним механизм оценки. Возьмем якорь QR и k.score наиболее похожих. Получим два множества клеток в образцах Query и Reference, состоящих из концов этих якорей. Найдем ближайших соседей клеток Q и R. Получится еще два множества. Чем больше одинаковых клеток в этих парах множеств, тем больший вес присваивается якорю QR. Другими словами, якоря должны связывать множества похожих клеток в одном образце с множеством похожих клеток в другом образце [164].

Далее выполняется непосредственно интеграция, а именно, объединение пар клеток, помеченных как один якорь с учетом его веса.

Таблица 5. Функции используемые в Seurat

Function	**Arguments**
FindIntegrationAnchors	normalization.method = "SCT", "LogNormalize"
	reduction = "rpca", "cca", "rlsi"
	nn.method = "rann", "annoy"
IntegrateData	normalization.method = "SCT", "LogNormalize"
	weight.reduction = string, vector of strings, vector of DimReduc objects, NULL

Для интеграции наборов данных scRNA-seq в алгоритм были внесены небольшие изменения. Вместо использования традиционного корреляционного анализа CCA для определения якорей используется рандомизированный нелинейный PCA анализ RPCA (Рис. 10). При поиске якорей между любыми двумя наборами данных с помощью RPCA каждый набор данных проецируется в пространство других главных компонент и отбираются якоря, соответствующие ранее установленным требованиям взаимного соседства. Команды для обоих алгоритмов во многом похожи, но эти два метода могут применяться в разных ситуациях.

Для определения общих источников вариаций между наборами данных хорошо подходит CCA. В данном случае клеточные типы известны и не меняются, а в паттернах экспрессий генов будут наблюдаться существенная разница между различными экспериментальными условиями. То есть CCA используется для интеграции, когда условия эксперимента или внешние воздействия вызывают очень сильные сдвиги экспрессии, при интеграции датасетов, содержащих различные типы данных (протеом, транскриптом, метаболом и т.д.) или полученные от организмов разных видов. Однако интеграция на основе CCA также может привести к чрезмерной коррекции, особенно когда большая часть клеток не перекрывается между датасетами.

Таким образом, не рекомендуется использовать метод интеграции по умолчанию, поскольку он может не соответствовать особенностям исследуемых клеток. На данном этапе очень важно взаимодействие биоинформатиков и биологов для выбора подходящего метода. Вместе с тем, не рекомендуется ограничиваться только одним выбранным методом, так как даже очень хорошее знание объекта исследование не может гарантировать исключительно правильного проведения всех этапов подготовки к секвенированию. Биологические объекты обладают крайне вариационными набором свойств, что оправдывает проведение расчетов с использованием различных методик.

Интеграция на основе RPCA выполняется значительно быстрее, а также представляет собой более консервативный подход, при котором клетки в разных биологических состояниях с меньшей вероятностью «выравниваются» после интеграции. Метод называется «sctransform» и позволяет избежать некоторых ошибок стандартных рабочих процессов нормализации, включая добавление псевдосчетчика и логарифмическое преобразование. sctransform выполняет нормализацию, стабилизацию дисперсии и выбор признаков на основе матрицы экспрессии генов на основе UMI [165]. Точно так же расширение sctransform может выполнять дифференциальное выражение непосредственно для результирующих оценок параметров вместо остаточных значений, потенциально связывая их с эмпирической байесовской структурой.

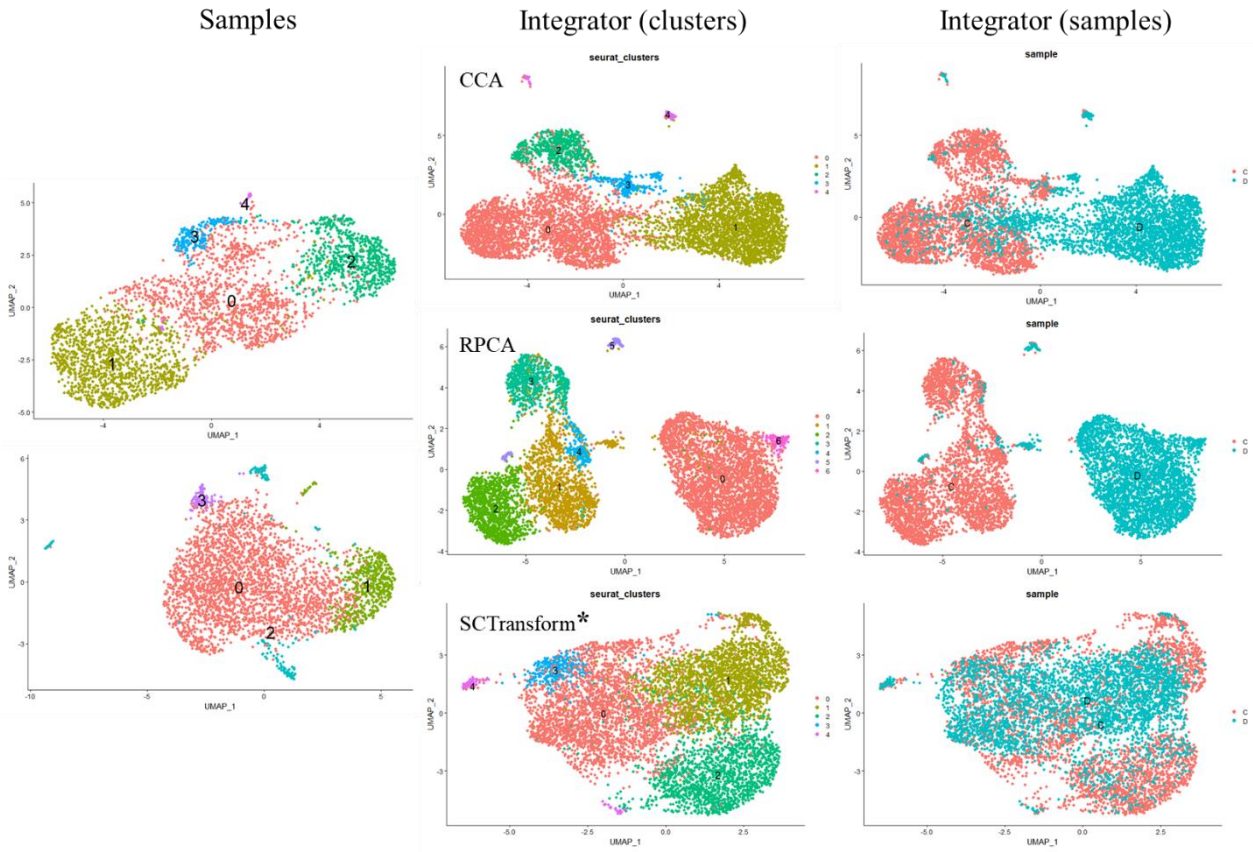


Рисунок 10. Кластеризация и различные методы интеграции в пакете Seurat (CCA - canonical correlation analysis, RPCA - reciprocal PCA, *SCTransform - регуляризованная отрицательная биномиальная регрессия (не является методом интеграции, используется на этапе нормализации, заменяет функции NormalizeData, FindVariableFeatures, ScaleData)

Интеграция, как один их этапов пайплайна обработки данных scRNA-seq появился не сразу. Необходимость в таком этапе возникла в связи со спецификой подготовки, хранения и культивирования клеточного материала. Далеко не всегда удается создать дизайн эксперимента, учитывающий все организационные вопросы, касающиеся забора клеточного материала, выделения из него клеток и т.д. Также, в силу ограниченных ресурсов при выполнении задач исследований ученые вынуждены прибегать к подходам, которые позволяют сэкономить время и средства. В результате клетки могут быть выделены в разное время, разными людьми, хранение, культивирование и подготовка библиотек для секвенирования, также зачастую выполняется впервые, без наличия надлежащего опыта. Все эти артефакты будут хорошо видны при интеграции. Интеграция ставит своей целью создать «интегрированный» анализ данных для последующего анализа, определить типы клеток, которые присутствуют в обоих наборах данных, получить маркеры типов клеток, которые сохраняются как в контрольных, так и в стимулированных клетках, сравнить наборы данных, чтобы найти специфические ответы типа клеток на стимуляцию. Помимо этого, есть и менее распространенная цель – увеличение количества клеток в образце или увеличение количества прочтений. Все это позволяет создавать интеграторы, анализируя которые можно выявить пропущенные закономерности при анализе массивов отдельного образца. Выбор алгоритма интеграции является крайне ответственным этапом, так как результат последующего downstream analysis будет зависеть именно от этого этапа.

Появившиеся в самом начале развития направления scRNA-seq алгоритмы анализа существенно эволюционировали не только за счет адаптации и модификации существующих математических подходов для анализа подобных видов данных, но и создания новых, которые заметно расширили современный стандартный пайплайн обработки данных scRNA-seq. Тем не менее, в алгоритме существуют ключевые этапы, от принятия правильных решений на которых зависит успешность полученных

результатов и их интерпретация. Практика показывает, что специалисты, впервые столкнувшиеся с задачей анализа этого типа данных, допускают схожие ошибки. Задачей дальнейшего развития является адаптация математических основ и методов, используемых в анализе данных scRNA-seq для специалистов биологического профиля. Также, работа содержит практические рекомендации по проведению основных этапов анализа данных scRNA-seq в R-пакете Seurat. Концептуальным вопросом, затронутым в статье, является, на наш взгляд, слепое следование стандартным существующим методам анализа данных scRNA-seq. Для анализа используются математические методы, используемые наравне с биологией и во многих других отраслях науки. Однако, сложность биологических данных заставляет задуматься над созданием собственных математических методов, учитывающих особенности и сложность биологических систем.

Разработать новый биоцентрический аналитический аппарат могут помочь такие новые направления в вычислительной математике, как системная и алгебраическая биология, теория систем и искусственный интеллект. Все эти направления помогут исследователям математически описать биологические процессы с активным участием естественнонаучных специалистов, разработать интегральные уравнения, учитывающие не только самые значимые переменные.

1.6.11 Типирование клеток

Последние достижения в области scRNA-seq позволили достичь высокого уровня детализации в характеристике изменений экспрессии генов. Методологии множественного анализа отдельных клеток были разработаны для обнаружения изменений экспрессии генов и кластеризации клеток по сходству экспрессии генов.

Однако классификация кластеров по типу клеток в значительной степени зависит от известных маркерных генов, а аннотация кластеров выполняется вручную. Эта стратегия страдает субъективностью и

ограничивает адекватную дифференциацию близкородственных подмножеств клеток. SingleR [166], новый вычислительный метод для объективного распознавание типа клеток scRNA-seq. SingleR использует эталонные транскриптомные наборы данных чистых типов клеток пакета celldex для определения клетки предшественницы каждой отдельной клетки независимо.

Для вычислений с помощью функции BlueprintEncodeData используются нормализованные данные экспрессии 259 образцов bulk RNA-seq полученные Blueprint и ENCODE из популяций стромальных и иммунных клеток [167, 168]. Данные были обработаны, нормализованы и описаны [169], сырые данные были загружены из Blueprint и ENCODE в 2016 году и нормализованы с помощью edgeR (TPMs).

Blueprint Epigenomics содержит 144 чистых образца RNA-seq с 28 аннотированными типами клеток. Encode содержит 115 RNA-seq с 17 аннотированными типами клеток. Вместе, референсы содержат 259 образца и 43 проаннотированных типов клеток.

Функция HumanPrimaryCellAtlasData дает доступ к 713 нормализованным микрочиповым образцам из Human Primary Cell Atlas (HPCA) [170]. Данные были обработаны, нормализованы и описаны, все образцы приведены к 37 основным клеточным типам и 157 подтипам.

Аннотации SingleR в сочетании с Seurat, пакет обработки и анализа, разработанный для scRNA-seq, представляет собой мощный инструмент для исследования данных scRNA-seq. Пакет R разработан для создания аннотированных объектов scRNA-seq, которые затем могут использовать веб-инструмент SingleR для визуализации и дальнейшего анализа данных.

Наряду со ставшим уже классическим алгоритмом анализа scRNA-seq данных – Cell Ranger, для более подробного анализа применяется R пакет Seurat, позволяющий корректировать ход анализа на каждом этапе.

На сегодняшний день не существует универсального биоинформатического инструмента для определения типов клеток в

результатах scRNA-seq. Для типирования обычно выявляют дифференциально экспрессирующиеся гены между кластерами и соотносят их с признанными маркерами различных типов клеток.

Для типирования клеток в полученных образцах был использован гибридный подход, заключающийся в применении нескольких методов определения типов клеток. В начале был проведен ряд глобальных (сравнение выбранной субпопуляции со всеми клетками библиотеки) и локальных (сравнение двух выбранных субпопуляций друг с другом) сравнений дифференциально экспрессирующихся генов в выделенных субпопуляциях. Затем был проведен контроль распределения в библиотеке маркеров интересующих нас типов клеток по клеточным референсам (PanglaoDB [171], CellMarker [172]), после этого использовали пакеты alona [173] (в основе пакет Python - adobo) и SingleR (Bioconductor) для автоматического распознавания типов клеток (на основании информации из клеточных референсов HumanPrimaryCellAtlasData [174] и BlueprintEncodeData [175]).

Задача типирования клеток на сегодняшний день одна из самых сложных и неоднозначных. Определяя тип клеток по специфическим маркерам, предполагается, что в образце присутствуют только конечно дифференцированные формы. На самом деле в образце могут быть лишь одна или две субпопуляции со специфическими маркерами. Остальные клетки находятся в промежуточных, переходных формах на пути дифференцировки и идентифицировать их по характерным маркерам не представляется возможным.

После получения кластеров клеток была рассчитана дифференциальная экспрессия генов в этих кластерах. Оказалось, что в каких-то кластерах можно довольно точно определить тип клеток по специфическим маркерам, указанным в базах данных (PanglaoDB [176], CellMarker [177]). В других же кластерах среди дифференциально экспрессирующихся генов не оказалось специфических маркеров, и было предположено, что можно типировать

переходные формы клеток по происходящим в них процессах. Для проверки этого предположения были использованы списки наиболее представленных в этих кластерах генов. Проведя кластеризацию по GO: Biological process с помощью он-лайн сервиса String [178] были охарактеризованы не типированные кластеры клеток по биологическим процессам.

Кроме стандартных методов типирования клеток можно использовать альтернативные. В данной работе предложено несколько таких подходов. Один из них – типирование по переходным генам (driver genes). Переходные гены были предложены в динамической модели процессинга scVelo. Задача решается в рамках модели максимизации ожидания, основанной на правдоподобии, путем итеративной оценки параметров скорости реакции и латентных переменных, специфичных для каждой клетки, т.е. транскрипционное состояние и внутреннее латентное время клетки. Тем самым задача заключается в получении фазовой траектории соотношений несплайсированных и сплайсированных форм для каждого гена.

С помощью heatmap построенной по таблице значений латентного времени можно выделить группы генов, уровень экспрессии которых увеличивается в определенные временные состояния жизни клетки. Такие переходные гены определяются по принципу высокой представленности в критических точках перехода клетки из одного состояния в другое, например, при дифференцировке или реакции на предъявляемые ей стимулы. Аннотация переходных генов может помочь понять за счет активации каких биологических процессов у клетки появилась возможность продвинуться в развитии, перейдя из одного состояния в другое (Рис. 11).

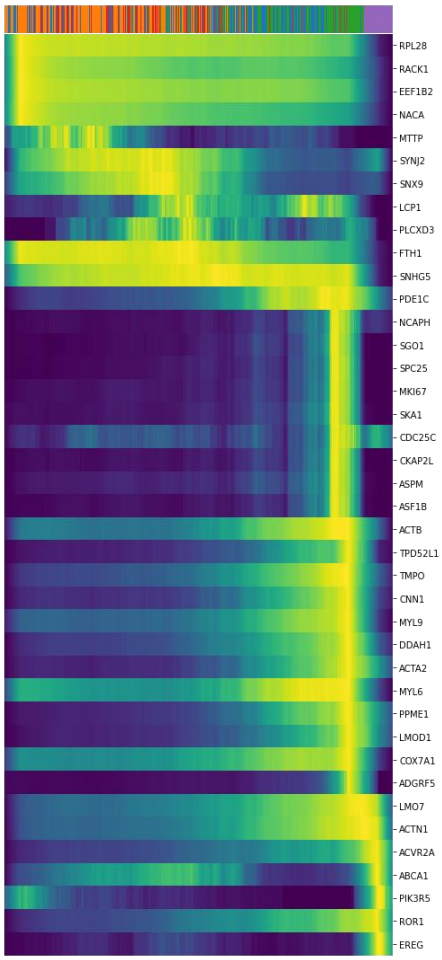


Рисунок 11. Гены, связанные с траекторией, классифицируются как ранние, переходные и поздние, в зависимости от времени их пика экспрессии. Ось X: гены упорядочены по времени их пика экспрессии, Ось Y: клетки упорядочены по псевдовремени. Цветная линия сверху тепловой карты соответствует цветам кластеров.

Динамика экспрессии, разрешенная в течение латентного времени, демонстрирует четкий каскад транскрипции в 300 лучших генах с рейтингом вероятности.

Скрытые временные точки, специфичные для генов, полученные из динамической модели, связаны с универсальным общим геном латентным временем, которое представляет внутренние часы клетки и основано только на ее транскрипционной динамике.

Этот метод подразумевает создание референса, состоящего из большого количества аннотированных вручную образцов. Затем этот референс используется для аннотирования новых исследуемых образцов с неизвестными типами клеток. В основе такого переноса лежат два основных

метода – перенос маркеров и проецирование [179]. Важно, что для аннотации новых данных нет необходимости в их предварительной предобработке.

Типирование клеток - очень ответственный этап в обработке данных scRNA-seq. Необходимость развития автоматических методов типирования объективно оправдан с точки зрения времязатратности этой задачи при ее реализации вручную. Помимо ручного типирования и автоматического типирования, основанного на переносе маркеров с уже аннотированных образцов на еще не аннотированные, стали развиваться методы, в которых активно используются технологии искусственного интеллекта – машинное обучение и нейросети.

Нейросеть scCapsNet-mask [180] основана на своей предыдущей версии scCapsNet и имеет два принципиальных отличия. Во-первых, в scCapsNet-mask есть I нейронных сетей, соответствующих I типам клеток в наборе данных, и каждая нейронная сеть использует блок линейной ректификации Rectified Linear Unit (ReLU) или tanh в качестве функции активации. Напротив, в scCapsNet количество нейронных сетей обозначается как гиперпараметр. Во-вторых, в процессе «динамической маршрутизации» у scCapsNet-mask есть дополнительный шаг после вычисления коэффициента связи. Коэффициенты связи необходимо поэлементно умножать на маску матрицы, которая является первичной матрицей, в которой все диагональные элементы равны единице, а недиагональные элементы равны нулю. Эта операция концентрирует веса в диагональных элементах и игнорирует недиагональные элементы коэффициента связи.

Метод scCapsnet_mask [181] представляет собой систему нейронных сетей (Рис. 12). На первом слое n полносвязных нейронных сетей отображают исходную матрицу подсчета в некоторые вектора (первичные капсулы). На следующих слоях вектора подаются в капсульную нейронную сеть, которая позволяет выявить вклад каждого вектора в капсулу типа.

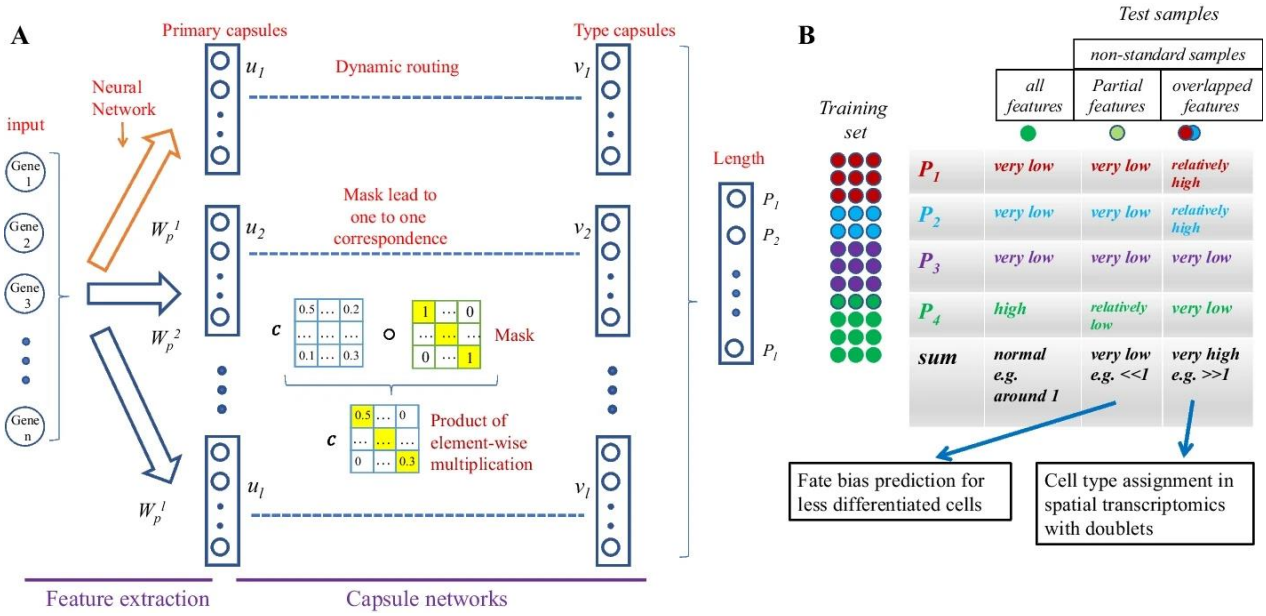


Рисунок 12. scCapsNet-mask [182] — это обновленная версия scCapsNet с возможностями упрощенной интерпретации модели и обработки нестандартных образцов

Input – входной слой; **Feature extraction** – выделение признаков; **Neural Network** – нейронная сеть; **Primary capsules** - первичные капсулы нейросети; **Dynamic routing** – динамический роут; **mask lead to one to one correspondence** – маска однозначного соответствия; **product of element-wise multiplication** – покомпонентное произведение; **Length** – длина; **Training set** – тренировочная выборка; **all features** – все свойства; **partial features** – частичные свойства; **overlapped features** – перекрывающиеся свойства; **fate bias predication for less differentiated cells** – предсказание смещения от направления развития клеток, находящихся в начальной стадии дифференцировки; **cell type assignment in spatial transcriptomic with doublets** – оценка клеточного типа в пространственной транскриптомике с дублетами (задвоенными клетками)

1.6.12 Траектории развития

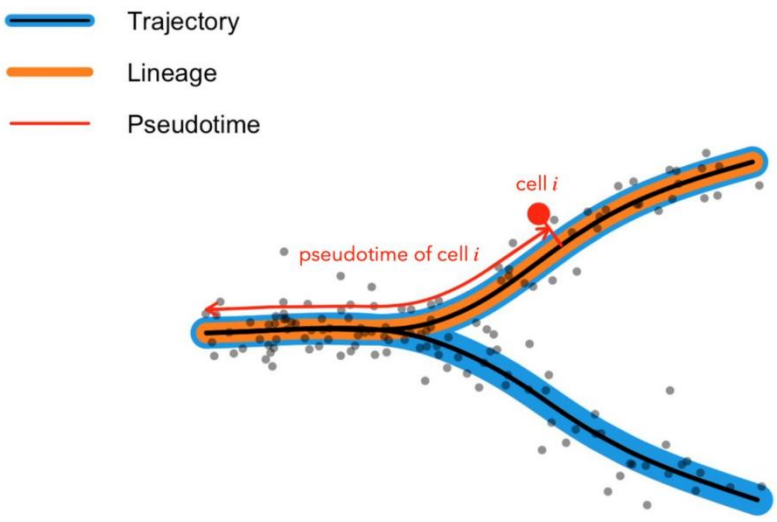


Рисунок 13. Основные термины, используемые в методе построения траекторий развития. Траектория представляет собой набор направлений развития (одной или нескольких). Направление развития представляет собой индивидуальный путь развития. Псевдовременем называют длину линии развития от ее формирования до проекции определенной клетки на данную линию [183]. **Trajectory** – траектория развития; **Lineage** – направление дифференцировки; **Pseudotime** - псевдовремя

Методы построения траектории оценивают траекторию, набор путей дифференцировки динамической системы, упорядочивая клетки по путям такого динамического процесса. В случае дифференцировки или любого процесса, в котором прогрессия происходит в одном направлении, значение псевдовремени клетки можно рассматривать как показатель того, насколько далеко в спектре находится клетка относительно состояния предшественника (Рис. 13) [184].

В методе построения траекторий развития используется понятия латентного времени и псевдовремени. **Латентное время** – позволяет более глубоко взглянуть на процессы, происходящие в клетке. Тут учитываются процессы процессинга незрелых мРНК в зрелые и учитывается интронная информация. **Псевдовремя** – основывается только на экзонной информации, поэтому построение траекторий развития в псевдовремени основывается только на схожести по уровню экспрессии генов, а не на их степени зрелости [185].

В то время, как традиционные методы построения траекторий располагают клетки вдоль линии псевдовремени, основываясь на клеточной динамике клеток, находящихся на разных стадиях развития изучаемой популяции RNA velocity основана на динамической модели, описывающей динамику сплайсинга [186]. Первое принципиально отличие между псевдовременем и латентным временем состоит в доступной на момент расчета информации. В скорость включается дополнительная информация об интронах для построения модели преобразования пре-мРНК в мРНК, определения сходства клеток в их динамике и по прокси-вероятным переходам, в дополнение к изменениям в экспрессии самого гена.

В стандартном псевдовремени есть только «экзонная» информация, нет информации об интронах или динамике их сплайсинга и возможно клеточные переходы и клеточный порядок по сходству экспрессии генов, но поскольку нет возможности моделировать направление движения с использованием информации о сплайсинге может быть трудно определить, какой конец траектории является «корнем», без ручного указания с использованием чего-то вроде известных маркерных генов.

Топология траекторий развития может быть различной – от линейной до очень сложных (Рис. 14). Для некоторых методов необходимо указывать изначальную информацию - начальные и конечные клетки. При построении траекторий развития в некоторых случаях появляются точки бифуркации, которые располагаются в тех местах, где антикоррелирующие дистанции между ветвями становятся коррелирующими.

Абстракция графа на основе разделов (PAGA) решает эти фундаментальные проблемы, создавая графоподобные карты клеток, которые сохраняют как непрерывную, так и несвязанную структуру данных при различных разрешениях [187].

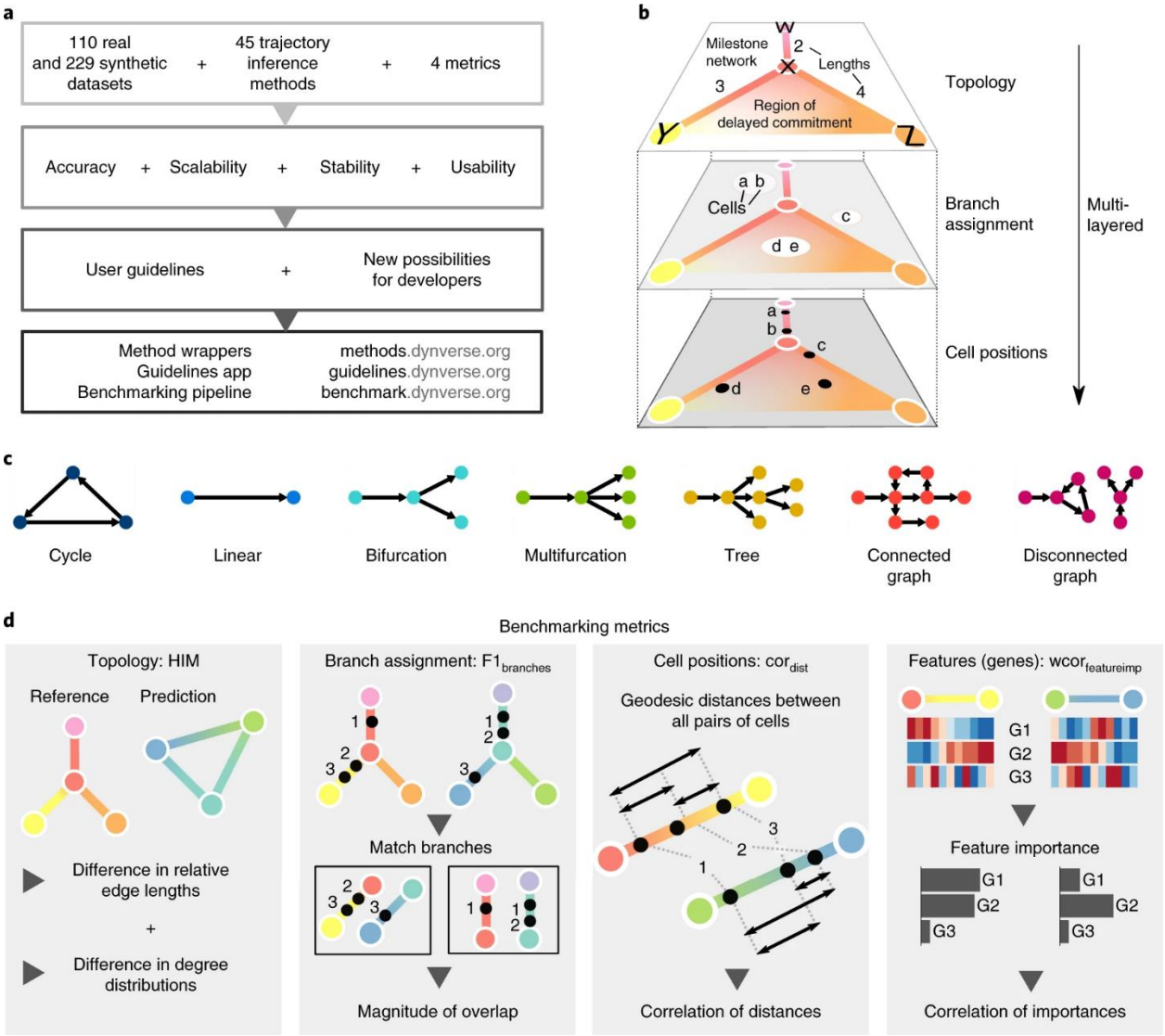


Рисунок 14. Обзор нескольких ключевых понятий построения траекторий [188]

110 real and 229 synthetic datasets – 110 оригинальных и 229 сгенерированных датасетов; **45 trajectory inference methods** – 45 методов построения траекторий развития; **4 metrics** – 4 метрики; **accuracy** - точность; **scalability** - масштабируемость; **stability** - стабильность; **usability** – удобство использования; **user guidelines** – инструкция пользователя; **new possibilities for developers** – новые возможности для разработчиков; **methods wrappers** – оберточный метод (метод оболочки); **guidelines app** - набор рекомендаций, правил; **benchmarking pipeline** – алгоритм изучения опыта конкурентов и внедрения лучших практик; **milestone network** – график опорных событий; **lengths** - длина; **topology** - топология; **branch assignment** – поиск точки ветвления; **cell positions** – положения клеток; **multi-layered** - многослойный; **cycle** – циклический граф; **linear** – линейный граф; **bifurcation** - ответвление; **tree** - граф, не имеющий простых циклов, называется лесом, связный граф, не имеющий простых циклов, называется деревом; **connected graph** – связный граф; **disconnected graph** – несвязный граф; **difference in relative edge lengths** - разница в относительных длинах ребер; **difference in degree distributions** - разница в распределении степеней; **match branches** – совпадение ответвлений; **magnitude of overlap** - величина перекрытия; **geodesic distance between all pairs of cells** - геодезическое расстояние между всеми парами клеток; **correlation of distances** – корреляция расстояний; **correlation of importances** – корреляция между важностями признаков

Slingshot сочетает в себе высокостабильные методы, необходимые для зашумленных данных по одной клетке, с гибкостью для идентификации нескольких родословных с различными уровнями контроля. Slingshot состоит из двух основных этапов: 1) вывод глобальной структуры происхождения и 2) вывод переменных псевдовремени для клеток вдоль каждого происхождения [189].

Monocle [190, 191, 192] выполняет дифференциальную экспрессию и анализ временных рядов для scRNA-seq экспериментов. Он упорядочивает отдельные клетки в соответствии с прогрессом в биологическом процессе, не зная заранее, какие гены определяют прогресс в этом процессе. Monocle также выполняет дифференциальный анализ экспрессии, кластеризацию, визуализацию и другие полезные задачи с данными экспрессии отдельных клеток. Он разработан для работы с данными RNA-Seq и qPCR, но может использоваться и с другими типами.

Исследования экспрессии генов отдельных клеток позволяют профилировать регуляцию транскрипции во время сложных биологических процессов и в очень гетерогенных популяциях клеток. Эти исследования позволяют обнаружить гены, которые идентифицируют определенные подтипы клеток, или которые отмечают определенные промежуточные состояния во время биологического процесса. Во многих scRNA-seq исследованиях, отдельные клетки выполняются через программу экспрессии генов несинхронизированным образом. В результате, каждая клетка представляет собой снимок исследуемой транскрипционной программы. Пакет monocle предоставляет инструменты для анализа экспериментов по экспрессии отдельных клеток. Monocle предлагает стратегию упорядочивания отдельных клеток в псевдовремя, размещая их по траектории, соответствующей биологическому процессу, например дифференцировке клеток. Monocle выстраивает эту траектория прямо из данных, либо полностью без контроля, либо под контролем. Он также выполняет дифференциальную экспрессию генов и кластеризацию для

определения важных генов и состояний клеток. Он разработан для RNA-Seq исследования, но может использоваться с другими анализами.

Другие методы построения траекторий развития и псевдовремен клеток основаны на различных методах уменьшения размерности, таких как карта диффузии (Diffusion maps), метод карт собственных значений лапласиана (Laplacian eigenmaps), анализ независимых компонентов (ICA), обратимое вложение графа (Reversed graph embedding), анализ главных компонент (PCA) и т.д.

Slingshot новый метод получения траекторий развития клеток и их псевдовремен из данных scRNA-seq. Slingshot может использовать любой из перечисленных выше методов снижения размерности.

Slingshot сочетает в себе высокостабильные методы, необходимые для шумных scRNA-seq данных с гибкостью для идентификации нескольких направлений развития с разным уровнем курируемости. Slingshot состоит из двух основных этапов: 1) получение общей картины закономерности развития клеток образца 2) получение псевдовременных переменных для клеток вдоль каждого направления развития. Как и другие методы, Slingshot использует кластерный MST для стабильной идентификации ключевых элементов общего направления развития клеток, т. е. количество направлений и места их разветвления. Это позволяет нам выявить новые направления развития, а также использовать предметно-ориентированные знания для наблюдения за частями дерева (например, терминальные клеточные состояния). Для второго этапа предложен новый метод, называемый одновременные основным кривые, чтобы подогнать плавные кривые ветвления к этим линиям тем самым транслируя знание общего направления развития в стабильные оценки лежащих на клеточном уровне псевдовременных переменных для каждой линии.

В дополнение к основному методологическому ядру Slingshot, компоненты, описанные выше для направления развития и псевдовремени, необходимо отметить важность выбора предшествующего анализа.

Действительно, большинство методов реконструкции псевдовремени будут явно или неявно требовать выбора определенных параметров на предыдущих этапах рабочего процесса. Например, снижение размерности помогает уменьшить количество шума в данных и в визуализации, но разнообразие доступных подходов обладает большим влиянием на конечный результат. Monocle рекомендует ICA или DDRTree, Waterfall и TSCAN анализ основных компонент (PCA), Embedder использует Лапласовы собственные карты, а Wishbone использует карты диффузии для анализа и t-SNE для визуализации. Учитывая большое разнообразие данных, генерируемых методами анализа отдельных клеток, маловероятно, что существует универсальное решение проблемы уменьшения размерности, нормализации и кластеризации. Эти этапы анализа очень важны, и из-за того, что различные методы построения траекторий развития требуют различных данных в начале анализа, их может быть сложно сравнивать. Slingshot не уточняет упомянутых выше методов, создан с гибкостью и модульностью, для более легкой интеграции, нормализации, уменьшения размерности и кластеризации, которые считаются наиболее подходящими для конкретного набора данных. Рекомендуемый процессинг данных scRNA-seq реализован в Rpackages Bioconductor. Конвейер включает в себя адаптивный к данным выбор процедуры нормализации (пакет scone), уменьшение размерности с помощью отрицательной биномиальной модели с нулевым раздуванием (пакет zinbwave), а также последовательного ансамблевого кластера на основе передискретизации (RSEC; пакет clusterExperiment).

Рисунок 15. Характеристика 45 методов построения траекторий развития[39]

Dynverse, представляет собой коллекцию R-пакетов для построения траекторий развития. Идея этой коллекции заключается в объединении всех известных на сегодняшний день методов построения траекторий в одной платформе, которой можно пользоваться онлайн либо в установочной версии. Увеличивающееся количество методов построения траекторий развития

привело к необходимости классифицировать их по подходам к снижению размерности, необходимости указания исходных данных, математической основе. Коллекция Dynverse удобна концепцией «все в одном». Загрузив один раз данные у пользователя есть возможность получить траектории развития различными методами и выбрать наиболее оптимальный, наилучшим образом описывающий биологические процессы, происходящие в клетках образца (Рис. 15). Для получения траекторий развития онлайн, часто используется Asc-Seurat [193] (веб-сервис для анализа данных scRNA-seq) на основе Shiny [194]. Сервис включает в себя возможности известных библиотек Seurat [195] и Dynverse [196] а также функциональную аннотацию генов интереса BioMart [197].

Monocle [198, 199, 200] выполняет дифференциальную экспрессию и анализ временных рядов для экспериментов по экспрессии отдельных клеток. Он упорядочивает отдельные клетки в соответствии с прогрессом в биологическом процессе, не зная заранее, какие гены определяют прогресс в этом процессе. Monocle также выполняет дифференциальный анализ экспрессии, кластеризацию, визуализацию и другие полезные задачи с данными экспрессии отдельных клеток. Он разработан для работы с данными RNA-Seq и qPCR, но может использоваться и с другими типами.

Исследования экспрессии генов отдельных клеток позволяют профилировать регуляцию транскрипции во время сложных биологических процессов и в очень гетерогенных популяциях клеток. Эти исследования позволяют обнаружить гены, которые идентифицируют определенные подтипы клеток, или которые отмечают определенные промежуточные состояния во время биологического процесса. Во многих scRNA-seq исследования, отдельные клетки выполняются через программу экспрессии генов несинхронизированным образом. В результате, каждая клетка представляет собой снимок исследуемой транскрипционной программы. Пакет monocle предоставляет инструменты для анализа экспериментов по экспрессии отдельных клеток. Monocle представил стратегию

упорядочивания отдельных клеток в псевдовремя, размещая их по траектории, соответствующей биологическому процессу, например дифференцировке клеток. Monocle выстраивает эту траектория прямо из данных, либо полностью без присмотра, либо под контролем. Он также выполняет дифференциальная экспрессия генов и кластеризация для определения важных генов и состояний клеток. Он разработан для RNA-Seq исследования, но может использоваться с другими анализами.

Другие методы построения траекторий развития и псевдовремен клеток основаны на различных методах уменьшения размерности, таких как карта диффузии (Diffusion maps), метод карт собственных значений лапласиана (Laplacian eigenmaps), анализ независимых компонентов (ICA), обратимое вложение графа (Reversed graph embedding), анализ главных компонент (PCA) и т.д.

Slingshot [201] новый метод получения траекторий развития клеток и их псевдовремен из данных scRNA-seq. Slingshot может использовать любой из перечисленных выше методов снижения размерности.

Slingshot сочетает в себе высокостабильные методы, необходимые для шумных scRNA-seq данных с гибкостью для идентификации нескольких направлений развития с разным уровнем курируемости. Slingshot состоит из двух основных этапов: 1) получение общей картины закономерности развития клеток образца 2) получение псевдовременных переменных для клеток вдоль каждого направления развития. Как и другие методы, Slingshot использует кластерный MST для стабильной идентификации ключевых элементов общего направления развития клеток, т. е. количество направлений и места их разветвления. Это позволяет нам выявить новые направления развития, а также использовать предметно-ориентированные знания для наблюдения за частями дерева (например, терминальные клеточные состояния). Для второго этапа предложен новый метод, называемый одновременные основным кривые, чтобы подогнать плавные кривые ветвления к этим линиям тем самым транслируя знание общего

направления развития в стабильные оценки лежащих на клеточном уровне псевдовременных переменных для каждой линии.

В дополнение к основному методологическому ядру Slingshot, компоненты, описанные выше для направления развития и псевдовремени, необходимо отметить важность выбора предшествующего анализа. Действительно, большинство методов реконструкции псевдовремени будут явно или неявно требовать выбора определенных параметров на предыдущих этапах рабочего процесса. Например, снижение размерности помогает уменьшить количество шума в данных и в визуализации, но разнообразие доступных подходов обладает большим влиянием на конечный результат. Monocle рекомендует ICA или DDRTree, Waterfall и TSCAN анализ основных компонент (PCA), Embedder использует Лапласовы собственные карты, а Wishbone использует карты диффузии для анализа и t-SNE для визуализации. Учитывая большое разнообразие данных, генерируемых методами анализа отдельных клеток, маловероятно, что существует универсальное решение проблемы уменьшения размерности, нормализации и кластеризации. Эти этапы анализа очень важны, и из-за того, что различные методы построения траекторий развития требуют различных данных в начале анализа, их может быть сложно сравнивать. Slingshot не уточняет упомянутых выше методов, создан с гибкостью и модульностью, для более легкой интеграции, нормализации, уменьшения размерности и кластеризации, которые считаются наиболее подходящими для конкретного набора данных. Рекомендуемый процессинг данных scRNA-seq реализован в Rpackages Bioconductor [202]. Конвейер включает в себя адаптивный к данным выбор процедуры нормализации (пакет scone [203]), уменьшение размерности с помощью отрицательная биномиальная модель с нулевым раздуванием (пакет zinbwave [204]), а также последовательного ансамблевого кластера на основе передискретизации (RSEC; пакет clusterExperiment [205]).

Dynverse [206, 207] представляет собой коллекцию R-пакетов для построения траекторий развития. Идея этой коллекции заключается в

объединении всех известных на сегодняшний день методов построения траекторий в одной платформе, которой можно пользоваться он-лайн либо в установочной версии. Увеличивающееся количество методов построения траекторий развития привело к необходимости классифицировать их по подходам к снижению размерности, необходимости указания исходных данных, математической основе. Коллекция Dynverse удобна концепцией «все в одном». Загрузив один раз данные у пользователя есть возможность получить траектории развития различными методами и выбрать наиболее оптимальный, наилучшим образом описывающий биологические процессы, происходящие в клетках образца.

На графиках видно уже не просто кластеры, а общие вектора, направления развития клеток, которые обычно называют траекториями развития. До получения таких графиков получают график с кластеризацией, где у каждой точки (клетки) есть свои координаты, соответствующие ее мере схожести с соседними точками (клетками). Но на таком графике кластеризации невозможно предположить направление дифференцировки или траекторию развития клеток.

Для этого к матрицам экспрессии, полученным в программном конвейере Cell Ranger из .gtf-файла добавляется информация о выравнивании транскрипта на участок генома. В зависимости от условий каждый транскрипт получает дополнительную характеристику – экзонный (сплайсированный), интронный (несплайсированный), неопределенный (который не картируется ни на экзонную ни на интронную часть).

После этого для каждой клетки образца строится индивидуальный вектор, характеризующий направление развития этой клетки. Вектора все вместе формируют общий вектор, направление дифференцировки или траекторию развития клеток.

Большая сложность состоит в интерпретации таких результатов, поскольку неизвестно заранее, какие типы клеток в каких кластерах сгруппированы. Поэтому, перед построением векторов кластеры клеток

необходимо типировать. Проблема состоит в том, что не се клетки в образце находятся в конечно дифференцированных формах и трудно их однозначно типировать по специфическим маркерам. Поэтому можно использовать типирование по биологическим процессам.

Также, помощь в типировании оказывает и сам метод построения векторов RNA-velocity. Если изучить получаемые графики видно, что некоторые из кластеров находятся на пути дифференцировки из одной точки в другую. Если предположить, что нам известен тип клеток в начальной точке и в конечной, но не известен тип клеток в промежуточном кластере, то тип клеток в промежуточном кластере можно определить исходя из литературных знаний об изменении транскриптома клеток в процессе дифференцировки. В том случае, если описанные биологические процессы, определенные в этом промежуточном кластере, совпадут с литературными данными, можно считать, что задача типирования клеток для этого кластера решена.

1.6.13 RNA-velocity

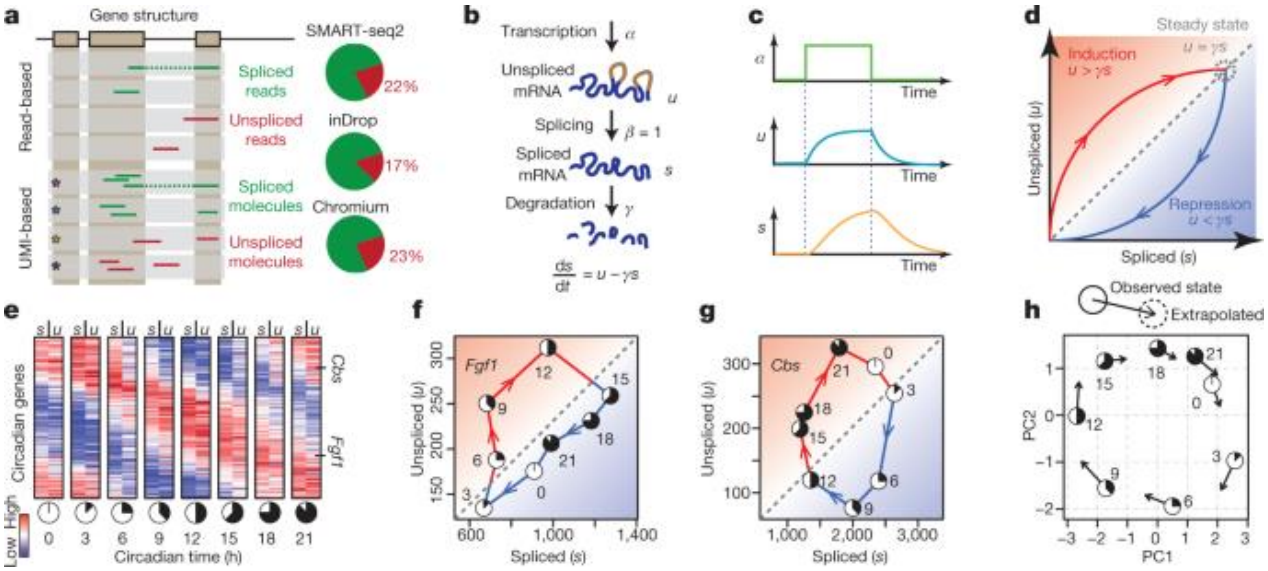


Рисунок 16. Предсказание клеточного состояния по соотношению несплайсированных и сплайсированных мРНК [208]

UMI-based – методы, основанные на уникальных молекулярных идентификаторах; **Read-based** – методы, основанные на прочтениях; **Gene structure** – структура гена; **Spliced/Unspliced** – после сплайсинга, до сплайсинга; **Steady state** – стационарное состояние; **Induction/Repression** –

Индукция(стимуляция)/Репрессия; **Circadian time** – циркадный ритм; **Observed/Extrapolated** – наблюдаемое/прогнозируемое

Большим недостатком существующих методов обработки данных scRNA-seq является отсутствие возможности наблюдать процессы во времени. Такие возможности были бы полезны в эмбриологии, дифференцировке и регенерации тканей. В 2018 году группа Манно, анализируя результаты секвенирования одиночных клеток, библиотеки для которого были подготовлены по протоколам SMART-seq2, STRT/C1, inDrop and 10x Genomics Chromium, обнаружили, что в 15-25% процентах прочтений содержатся несплайсированные интронные последовательности (Рис. 16 а).

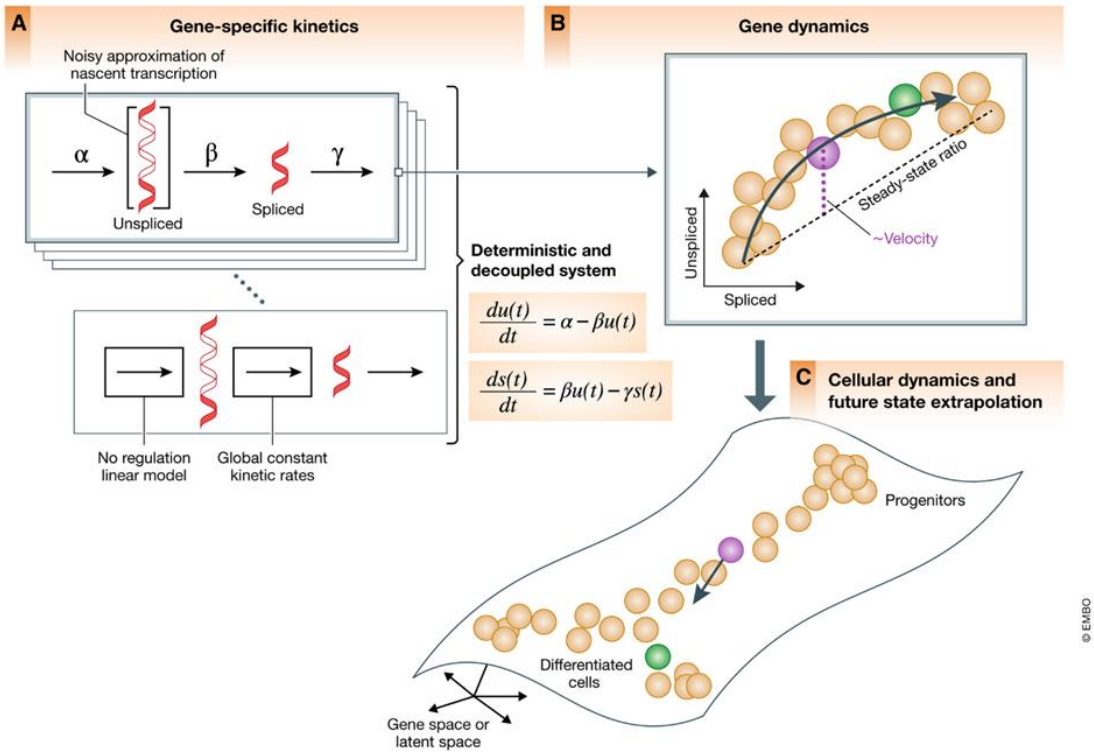


Рисунок 17. Учет кинетики сплайсинга используется при расчете RNA-velocity в переходных популяциях [209]

Gene-specific kinetics – кинетика на уровне гена; **noisy approximation of nascent transcription** – шумная аппроксимация начинающейся транскрипции; **deterministic and decoupled system** - ; **steady-state ratio** – соотношение в стационарном состоянии; **velocity** – скорость транскрипционной динамики; **cellular dynamics and future extrapolation** – клеточная динамика и предсказание состояний в ближайшем будущем; **Progenitors** – клетки-предшественницы; **differentiated cells** – дифференцированные клетки; **Gene space or latent space** – пространство генов или латентного времени

Для того, чтобы рассчитать зависимость от времени соотношения предшественников и зрелых форм мРНК была предложена простейшая

модель транскрипционной динамики, в которой первая производная по времени сплайсированных форм определялась как соотношение между сплайсированными и несплайсированными формами и деградацией [210] (Рис. 17).

Так, изначально существовали две модели – предложенная в начале “steadystate” модель **velocyto** и впоследствии расширенная динамическая модель **scVelo** (Рис. 18).

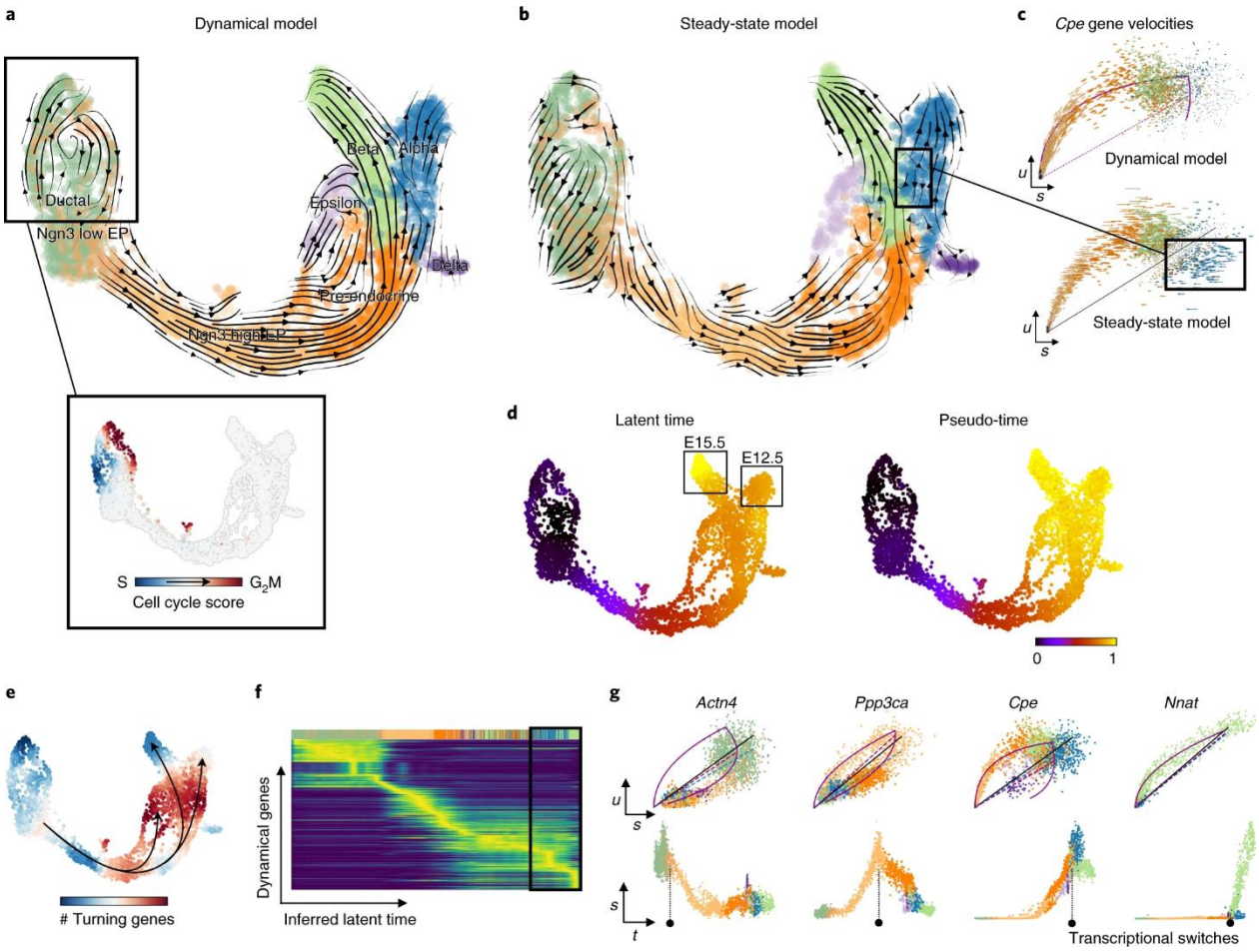


Рисунок 18. Выявление циклических популяций, детерминация направления клеточной линии дифференцировки (коммитирование), определение транскрипционной активности и судьбы клетки в псеводвремени

Dynamical model – динамическая модель; **steady-state model** – статичная модель; **cell cycle score** – оценка отношения клетки к определенной фазе клеточного цикла; **latent time** – латентное время (внутреннее время клетки, связано с показателями транскрипционной динамики); **pseudotime** – псевдовремя (время популяции, показывает, где находится клетка на линии, соединяющей наименее дифференцированную и наиболее дифференцированную клетки); **Turning genes** – активирующиеся гены; **dynamical genes** – динамические гены (как правило находятся в местах бифуркации траекторий развития, играя ключевую роль в определении клеточной судьбы)

Стационарная модель оценивает скорости как отклонение наблюдаемого соотношения несплайсированных и сплайсированных мРНК от предполагаемого стационарного отношения. Стационарное отношение аппроксимируется линейной регрессией для клеток, находящихся в нижнем и верхнем квантилях и достигших стационарного уровня экспрессии. Эта модель основана на двух основных предположения: общая скорость сплайсинга генов и ограниченное наблюдение за стационарными уровнями экспрессии в исследуемых данных. Хотя эти допущения обеспечивают надежную оценку, они могут привести к ошибкам в оценках скорости и клеточных состояний при их нарушении из-за неоднородности субпопуляции или невозможности наблюдения за системой вблизи ее устойчивого состояния.

Недавно представленная **динамическая модель**, основанная на правдоподобии, обобщает оценку скорости в переходных системах. Несмотря на ослаблении предположения об установившемся режиме, кинетика объясняется детерминированной и полностью несвязанной системой линейных дифференциальных уравнений с постоянными кинетическими параметрами скорости.

Позже, по мере развития идеи, в 2021 появились предложения по дальнейшему развитию метода. Основное предложение заключалось в усложнении существующих моделей. Были предложены следующие модели [211]:

- модель с более сложной кинетикой, учитывающей динамику генов (Рис. 19 а)

Предыдущие модели были основаны на предположениях, которые не учитывали кинетических вариаций в транскрипционных модуляциях, сплайсинге и уровне деградации на более сложном уровне. Транскрипционная динамика существующих моделей описывалась простым кинетическим уравнением первого порядка. Такому уравнению подчиняется

лишь небольшое количество генов. Развитие данной модели сделает возможным классификацию клеток по их кинетическим режимам.

- **стохастическая модель, модель регуляции гена** (Рис. 19 b)

В отличие от стационарной модели учитывает не только соотношение между сплайсированными и несплайсированными формами, но и другие ковариации.

Модель динамической экспрессии генов может быть расширена до многомерной модели, описывающей не только переходы между состояниями клетки, но и регуляторные взаимодействия на этих переходах. Технологические достижения и включение новых функциональных слоев генома, таких как фактор транскрипции связывание, мотивы регуляторной последовательности, модификации хроматина и посредники, такие как активность РНК-полимеразы, имеют большие перспективы. Эти дополнительные показания предоставят информативные априорные данные о регуляторных сетях и расширенные спецификации кинетических моделей.

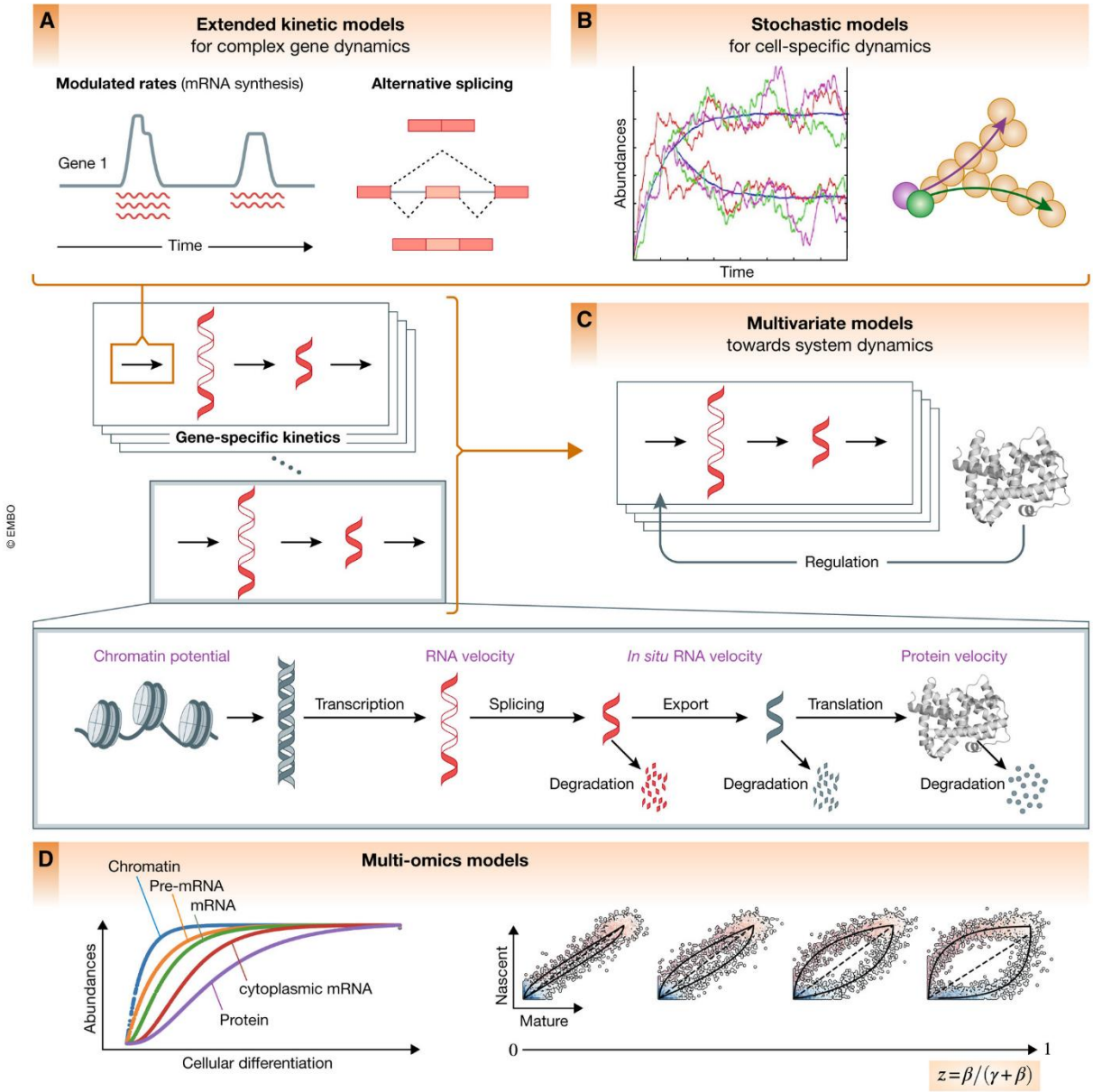


Рисунок 19. Концептуальные будущие направления модификации существующей модели транскрипционной динамики [212]

Extended kinetic models for complex gene dynamics – модель с более сложной кинетикой, учитывающей динамику генов; **stochastic models** - стохастическая модель, модель регуляции гена; **multivariate models** - мультивариативная модель; **multi-omics models** - с использованием омиксных данных

- мультивариативная модель (Рис. 19 c)

Модель динамической экспрессии генов может быть расширена до многомерного модель, описывающая не только переходы между состояниями клетки, но и регуляторные взаимодействия на этих переходах. Регуляторные мероприятия могут быть статистически наблюдаемые изменения экспрессии в течение псевдовремени. К описывают эти события, паттерны экспрессии генов-мишеней могут быть моделируется как функция активности факторов

транскрипции, в идеале рассматривается как нелинейная система, например, с использованием кинетики Хилла.

- с использованием омиксных данных (Рис. 19 d)

Скорость РНК основана на соединении измерений с лежащим в основе механизм (сплайсинг мРНК) с двумя модальностями, представляющими текущее и будущее состояние. В дополнение к экзонам и интронам сигналы, другие омики и молекулярные фрагменты могут быть использованы, если такие измерения доступны в непредвзятом режиме [213]. Изучение других модальностей имеет решающее значение для систем, где транскрипционная динамика сплайсинга мРНК не обеспечивает достаточного сигнала, например, если скорость сплайсинга относительно небольшая, в отличие от большой скорости деградации. Проблема недостаточного сигнала представляет собой проблему для текущей модели сплайсинга мРНК, но может быть разрешена, например, путем анализа других модальностей, например, с использованием динамики белков, где можно было бы ожидать увеличение кинетической характеристики статистической мощности с 0,5 до 0,8, при пятикратном увеличении период полураспада в белках в отличие от РНК. Для таких фрагментов как кэпированные, полиаденилированные и деградированные фрагменты транскриптов или избыток белка, расширение модели является простым после пересмотра основных предположений и статистических моделей специфичных для группы фрагментов, обеспечивая надежную количественную оценку.

Алгоритм Velocyto [214] в процессе работы создает несколько матриц, в которых учитывается не только картирование на экзонные участки генома, но также на интронные и промежуточные. В результате работы этого алгоритма формируется .loom-файл, содержащий информацию о количестве сплайсированных, несплайсированных и неоднозначных транскриптах.

Для аннотирования и отнесения прочтений к одной из трех групп, используются следующие правила:

Молекула аннотируется как **сплайсированная**, если все чтения в наборе, поддерживающем данную молекулу, выравниваются только на экзонные области совместимых транскриптов.

Молекула аннотируется как **несплайсированная**, если все совместимые модели транскриптов имели хотя бы одно чтение среди поддерживающего набора чтений для этого картирования молекулы, которое охватывает границу экзон-интрон, или выравнивается на интрон этого транскрипта.

Молекулы, для которых некоторые из совместимых моделей транскриптов есть выравнивание только на экзон, в то время как другие включали интронные выравнивания, аннотируются как **неоднозначные** и не используются в последующем анализе.

Для получения .loom-файла в случае работы с 10x Chromium используется папка с вложенными подпапками outs, outs/analys and outs/filtered_gene_bc_matrices. Также необходим файл .gtf. gff/gtf - General Feature Format/General Transfer Format, текстовый формат, используемый для описания генов, повторов и других характеристик ДНК, РНК и белков, содержит девять обязательных полей. В целом, похож на .bed формат, но отличается порядком полей и более жесткой структурой.

Формат файлов .loom разработан для эффективного хранения больших наборов омиксных данных. Обычно такие данные имеют форму большой матрицы чисел вместе с метаданными для строк и столбцов. Например, данные scRNA-seq состоят из измерений экспрессии для всех генов (строк) в большом количестве клеток (столбцов), а также метаданных для генов (например, хромосомы, цепи, расположения, имени) и для клеток (например, виды, пол, штамм, GFP-положительные).

Формат .loom разработан для представления наборов данных таким образом, чтобы строки и столбцы обрабатывались одинаково. В случае необходимости группировки генов и клеток есть возможность применения PCA для обоих классов и проведение фильтрации на основе проверки

результатов качества. Базы данных SQL и другие решения для хранения данных почти всегда обрабатывают данные как таблицу, а не как матрицу, что очень затрудняет добавление произвольных метаданных в строки и столбцы. В формате .loom это реализовано очень просто. Несмотря на то, что наборы данных могут содержать десятки тысяч строк (генов) и сотни тысяч столбцов (клеток) формат .loom обеспечивает эффективный доступ к произвольным строками столбцам.

Формат аннотированной матрицы позволяет гибко представлять данные для любых задач и анализа. Например, результат кластеризации может быть сохранен как атрибут идентификатора кластера для каждой клетки. Результат использования методов снижения размерности, такие как PCA или t-SNE, аналогичным образом может быть сохранен в виде двух атрибутов, дающих координаты проекции каждой клетки.

Наконец, формат .loom-файла удобен для анализа наборов данных с использованием графиков. .loom поддерживает графики как строк (например, генов), так и столбцов (например, клеток), и в каждом файле может храниться несколько графиков.

Количество РНК это значимый индикатор состояния клетки. С помощью scRNA-seq появилась возможность изучать РНК клетки с высокой точностью, чувствительностью и производительностью. Однако, этот метод позволяет сделать лишь статический «снимок» транскриптома клетки в момент времени, что лишает нас возможности ответить на вопросы, связанные, например, с эмбриогенезом и регенерацией, так как это динамические процессы.

Попыткой изучить происходящие в клетке процессы не только в момент секвенирования является RNA velocity, производное от уровня экспрессии гена, оцениваемое разницей между сплайсированными и несплайсированными формами мРНК.

Относительное количество насцентной (несплайсированной) и зрелой (сплайсированной) мРНК может быть использовано для оценки уровня сплайсинга и деградации гена.

Интересное наблюдение сделали авторы RNA velocity. Изучая данные SMART-seq2, STRT/C1, inDrop и 10x Chromium они обнаружили, что 15-25% прочтений это несплайсированные интронные транскрипты.

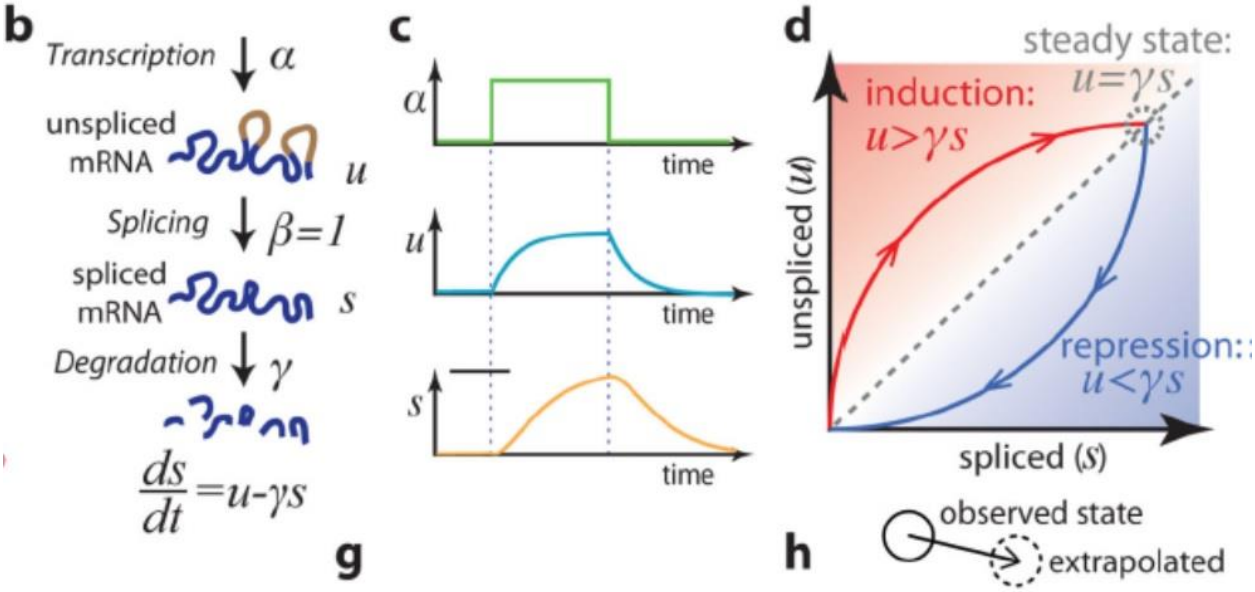


Рисунок 20. Биологическая основа RNA-velocity

Для квантификации количества предшественников и зрелой мРНК от времени была предложена упрощенная модель транскрипционной динамики, где первой производной по времени количества сплайсированной мРНК является соотношение между получением сплайсированной формы мРНК из несплайсированной и деградация мРНК. U (unspliced), s (spliced), α (transcription rate), u = γs, где γ – соотношение сплайсированной и деградированной, что соответствует соотношению длин интронных и экзонных участков (Рис. 20).

Для получения траекторий развития, графика псевдовремени, графиков соотношения сплайсированных и несплайсированных форм мРНК, а также графиков RNA-velocity и экспрессии генов используется пакет scVelo [215], который берет на вход .loom-файл, полученный с помощью Velocyto и .csv-файлы с выбранными для дальнейшего анализа кластерами.

Для измерения экспрессии генов в отдельных клетках их необходимо разрушить, после чего невозможно изучать динамику процессов в клетках и принятие клетками решений. Предложенный метод скорости РНК позволил восстановить направление транскрипционной динамики, используя тот факт, что недавно транскрибированные, несплайсированные пре-мРНК, обнаруживаемые по интронным участкам и зрелые, сплайсированные мРНК можно различить в протоколе scRNA-seq. Эта концепция измерения не только активности генов, но и их изменений в отдельных клетках (скорости РНК) открыла новые способы изучения клеточной дифференцировки. Первоначально предложенный алгоритм получает скорости как отклонение наблюдаемого соотношения сплайсированных и несплайсированных мРНК от предполагаемого устойчивого (стационарного) состояния. Ошибки в оценке скорости возникают, если основные допущения об общей скорости сплайсинга и наблюдении за полной динамикой сплайсинга с установившимися уровнями мРНК нарушаются.

С помощью scVelo эти ограничения устраняются путем анализа транскрипционной динамики кинетики сплайсинга с использованием динамической модели, основанной на правдоподобии. Это делает данный метод универсальным для целого ряда систем, включая переходные состояния клеток, которые являются обычными в эмбриологии развития и ответе на внешние стимулы. Кроме того, scVelo определяет специфичные для генов скорости транскрипции, сплайсинга и деградации и восстанавливает латентное время лежащих в основе клеточных процессов. Это скрытое время представляет собой внутренние часы, основанные на динамике транскрипции клетки и приблизительно соответствует реальному времени, в котором клетки дифференцируются. Более того, scVelo идентифицирует режимы регуляторных изменений, такие как стадии определения клеточной судьбы и выявляет предполагаемые гены-драйверы.

Выводы о направлениях траекторий строятся на использовании RNA velocity совместно с лежащей в основе кинетикой сплайсинга мРНК:

индукция транскрипции для конкретного гена приводит к увеличению (вновь транскрибируемых) несплайсированных мРНК предшественников, в то время как, наоборот, репрессия или отсутствие транскрипции приводит к уменьшению несплайсированных мРНК. Следовательно, отличая несплайсированной мРНК от сплайсированной, можно приблизительно оценить изменение количества мРНК (скорость РНК). Комбинация скоростей мРНК затем может быть использована для оценки будущего состояния отдельной клетки.

Оценка скорости РНК в настоящее время может быть решена с помощью трех существующих подходов:

стационарная / детерминированная модель (с использованием устойчивых остатков)

стохастическая модель (с использованием моментов второго порядка),

динамическая модель (с использованием вероятностной модели).

Стационарная / детерминированная модель, используемая в Velocyto, оценивает скорости следующим образом: в предположении, что фазы транскрипции (индукция и репрессия) длятся достаточно долго, чтобы достичь установившегося равновесия (активного и неактивного), скорости количественно оцениваются как отклонение наблюдаемого отношения от его стационарного отношения. Равновесные уровни мРНК аппроксимируются линейной регрессией по предполагаемым устойчивым состояниям в нижнем и верхнем квантилях. Это упрощение делает два фундаментальных предположения: общая скорость сплайсинга генов и стабильные уровни мРНК, которые должны быть отражены в данных. Это может привести к ошибкам в оценках скорости и клеточных состояний, поскольку предположения часто нарушаются, в частности, когда популяция включает в себя динамику нескольких гетерогенных субпопуляций.

Стохастическая модель нацелена на лучшее отображение устойчивых состояний. Рассматривая транскрипцию, сплайсинг и деградацию как вероятностные события, результирующий марковский процесс

аппроксимируется уравнениями моментов. Включая моменты второго порядка, он использует не только баланс несплайсированных и сплайсированных уровней мРНК, но и их ковариацию. На поджелудочной железе было продемонстрировано, что стохастичность добавляет ценную информацию, что в целом дает более высокую согласованность, чем детерминированная модель, при сохранении такой же эффективности во времени вычислений.

Динамическая модель (самая мощная и самая затратная с точки зрения вычислений) определяет всю динамику кинетики сплайсинга для каждого гена. Таким образом, он использует скорость РНК к широко варьирующимся характеристикам, таким как нестационарные популяции, поскольку не зависит от ограничений общей скорости сплайсинга или ранее описанного стационарного состояния.

Динамика сплайсинга решается в рамках вероятностной модели максимизации ожидания путем итеративной оценки параметров скорости реакции и скрытых переменных, специфичных для клетки, то есть состояния транскрипции k и внутреннего латентного времени t клетки.

Таким образом, он нацелен на изучение фазовой траектории несплайсированной/сплайсированной фазы. Моделируются четыре транскрипционных состояния для учета всех возможных конфигураций активности гена: два динамических переходных состояния (индукция и репрессия) и два устойчивых состояния (активное и неактивное), потенциально достигаемые после каждого динамического перехода.

На этапе ожидания для данной модельной оценки несплайсированной/ сплайсированной фазовой траектории латентное время присваивается наблюдаемому значению мРНК путем минимизации его расстояния до фазовой траектории. Затем транскрипционные состояния назначаются путем связывания вероятности с соответствующими сегментами на фазовой траектории (индукция, репрессия, активные и неактивные устойчивые

состояния). На этапе максимизации общая вероятность затем оптимизируется путем обновления параметров скорости реакции.

Модель дает более согласованные оценки скорости и лучшую идентификацию транскрипционных состояний. Кроме того, он позволяет систематически идентифицировать определяющие динамику гены на основе вероятности, тем самым находя ключевые факторы, которые управляют переходами клеточных судеб. Более того, динамическая модель предполагает универсальное внутреннее латентное время клетки, разделяемое между генами, что позволяет связывать гены и определять режимы транскрипционных изменений.

Динамическая модель восстанавливает скрытое время основных клеточных процессов. Это скрытое время представляет собой внутренние часы клетки и приблизительно соответствует реальному времени, в котором клетки дифференцируются, основываясь только на динамике транскрипции.

Латентное время scVelo основано только на его транскрипционной динамике и представляет внутренние часы клеток. Он улавливает аспекты фактического времени лучше, чем диффузное псевдовремени, основанное на сходстве. Латентное время позволяет временные отношения двух судеб, диффузия псевдовремени не различает их временное положение (Рис. 21).

Псевдовремя скорости - это измерение расстояния на графике скорости, основанное на случайном блуждании. После вычисления распределения по исходным клеткам, полученного из матрицы переходов с предполагаемой скоростью, он измеряет среднее число шагов, необходимых для достижения клетки после начала движения от одной из исходных клеток. В отличие от псевдовремени диффузии, он неявно определяет исходные клетки и основан на ориентированном графике скоростей, а не на основе подобия диффузионного ядра.

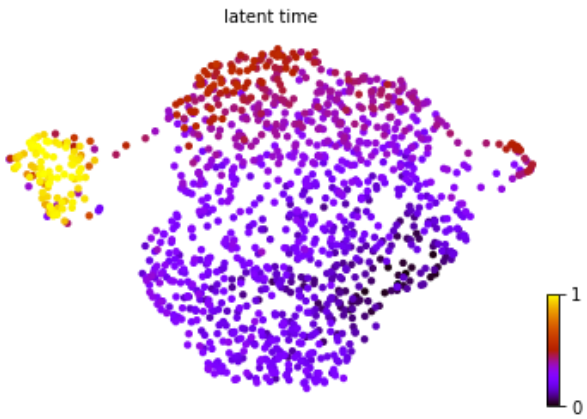


Рисунок 21. Латентное время scVelo

1.6.14 Анализ регулонов

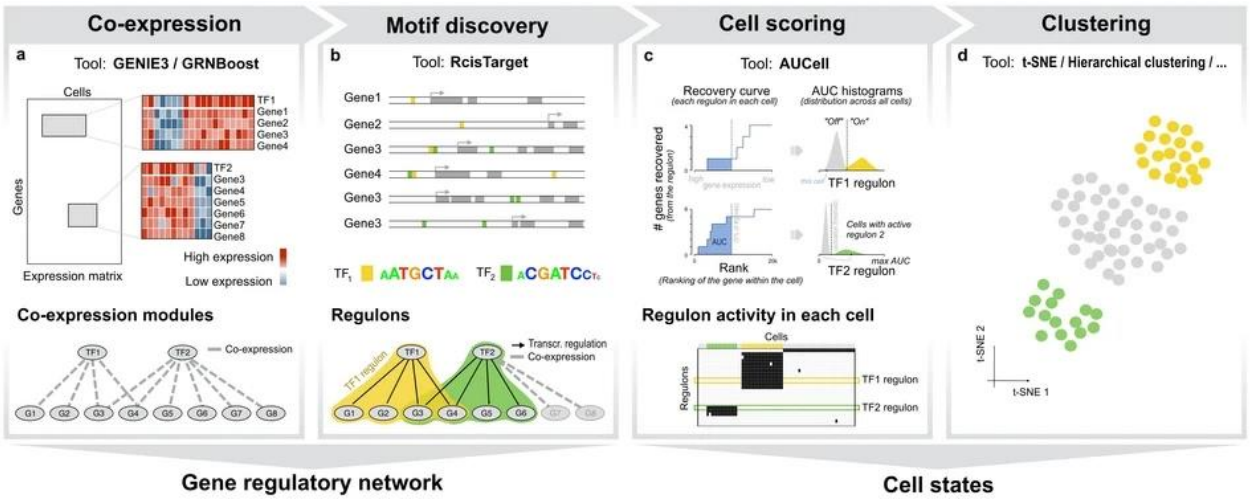


Рисунок 22. Рабочий алгоритм SCENIC

В 2017 году пайплайн scRNA-seq пополнился еще одним полезным инструментом – SCENIC [216]. Метод позволяет проанализировать транскрипционные факторы (ТФ), кофакторы и регулируемые ими гены (Рис. 22). В дополнение к аннотации по GO (Gene ontology) и RNA-velocity метод позволяет объяснить неочевидные, скрытые закономерности на уровне регуляции активности генов интереса. Алгоритм состоит из трех этапов. Сначала генерируются коэкпрессионные модули из ТФ и регулируемых ими генами-кандидатами. Затем с помощью пакета RcisTarget из генов-кандидатов в будущий регулон отбираются гены наиболее совпадающие по

мотивам связывания. Затем оценивается активность каждого из регулонов для каждой клетки, для объединения клеток в кластеры по данному признаку.

1.7 Современные исследования с использованием технологии 10x

Множество работ, ответов в которых нельзя было найти и которые не было смысла продолжать из-за отсутствия метода, позволяющего изучать каждую клетку образца получило дальнейшее развитие. Не имеющие объяснения явления стали более понятными и в связи с этим появились новые гипотезы о специфике ответа ткани на стимулы и о существовании, возможно, новых типов клеток, обладающих отличными свойствами.

Так, например в стромальной фракции подкожной жировой ткани были выявлены субпопуляции жировых стволовых клеток и клеток-предшественников. Были идентифицированы CD142 + клетки, которые способны регулировать адипогенез подавлением адипоцитов *in vivo* и *in vitro* паракринным образом [217].

Для выявления гетерогенности популяции стромально-васкулярной фракции клеток эпидидимальной (eWAT) и подвоздошной белой жировой ткани (iWAT) в контрольном образце и образце со стимуляцией адренергических рецепторов было проведено секвенирование 33000 клеток. Результат исследования показал, что в образцах эпидидимальной и подвздошной белой жировой ткани определились стволовые клетки адипоцитов, вступающие в процесс адипогенной дифференцировки по разному. Стимуляция рецепторов ADRB3 в eWAT вызывала значительное увеличение пролиферирующих стволовых клеток адипоцитов, дифференцировка которых в бежевые адипоциты начиналась из общей временной точки на траектории развития. С помощью секвенирования единичных клеток (scRNA-seq) были идентифицировал различные типы иммунных клеток в eWAT, включая субпопуляцию пролиферирующих макрофагов, которая занимает адипогенные ниши. Эти результаты демонстрируют способность scRNA-seq детально изучать адипогенные ниши

и предполагают новые функциональные взаимодействия между резидентными субпопуляциями стромальных клеток [218].

Секвенирование единичных клеток совместно с другими экспериментальными подходами было использовано для изучения определения судьбы клетки, взаимоотношений между клеточными поколениями, молекулярных детерминант на ранних стадиях сенсорного нейрогенеза у мышей. Полученные данные обеспечивают понимание структуры ландшафта Уодингтона последовательности клеточных поколений сенсорных клеток и предполагают, что взаимодействующие регуляторные сети генов активируются в незрелых постмитотических нейронах, приводят к появлению альтернативных транскрипционных программ последовательностей клеточных поколений, определяющих выбор судьбы нейрона [219].

Регуляция стволовых клеток жировой ткани (ASC) и адипогенез влияют на развитие метаболических осложнений, связанных с избытком жира в организме. Исследования на животных показали наличие различных субпопуляций ASC с разными способностями к дифференцировке. Использование ASC в терапии в качестве значимого источника мезенхимальных стволовых клеток требует детального изучения их свойств. Для того, чтобы охарактеризовать популяцию стволовых клеток в подкожной белой жировой ткани человека была использована транскриптомика одиночных клеток.

Транскриптом 574 клеток стромально-сосудистой фракции белой жировой ткани человека четырех здоровых женщин был проанализирован с помощью кластеризации и t-SNE визуализации. Затем идентифицированные популяции клеток были сопоставлены с типами клеток, присутствующими в WAT, с использованием данных микрочипа генов с экспрессионным профилем клеток стромально-васкулярной фракции, отсортированных с помощью проточной цитометрии. Клетки распределились в четыре отдельные кластера: три подтипа макрофагов, резидентных для жировой

ткани, и одна большая гомогенная популяция ASC. Несмотря на то, что при построении траекторий развития клеток в псевдовремени ASC находились на несколько разных стадиях дифференцировки, различия в экспрессии генов были небольшими и различить отдельные подтипы ASC было невозможно. По-видимому, у здоровых людей ASC составляют единую гомогенную клеточную популяцию, которую нельзя разделить с помощью транскриптомики единичных клеток, что указывает на общее происхождение адипоцитов человека в подкожной белой жировой ткани [220].

Мезенхимальные стволовые клетки, полученные из жировой ткани (ADSC), имеют большие перспективы для клинического применения в регенеративной медицине. В одном из исследований было проведено крупномасштабное одноклеточное транскриптомное секвенирование 24 358 культивируемых ADSC человека от трех доноров.

Клеточная гетерогенность является общей чертой биологических тканей и существует даже в пределах, казалось бы, «гомогенных» популяций стволовых клеток, на которые влияют внешние факторы микроокружения или внутренние факторы. Многочисленные доказательства подтверждают, что МСК в культуре по своей природе гетерогенны по фенотипам и функциям. Однако межклеточная изменчивость в культивируемой популяции МСК не может быть полностью описана с помощью нескольких маркеров клеточной поверхности. Отсутствие полного понимания клеточной гетерогенности МСК препятствовало разработке эффективного и воспроизводимого клинического применения.

В исследованиях транскриптомных данных единичных клеток влияние клеточного цикла часто рассматривается как биологический шум и исключается из данных.

Было обнаружено, что субпопуляции, идентифицированные путем кластеризации, обычно соответствуют клеткам, предположительно находящимся в одной и той же фазе клеточного цикла: 91,6% клеток в первой субпопуляции находились в фазе G1; 84,7% клеток в третьей субпопуляции

также находились в фазе G1; 68,8% клеток во 2 субпопуляции находились в S-фазе; 99,6% клеток в четвертой субпопуляции были идентифицированы как клетки фазы G2/M; и 59,1% клеток в 5 субпопуляции были идентифицированы как клетки S-фазы. Клетки, экспрессирующие характерные гены одной и той же фазы клеточного цикла, как правило, группируются вместе, как показано на примере распределения интенсивности экспрессии маркерных генов S-фазы (PCNA, MCM5), маркерных генов фазы G2/M (CCNF, CENPF), которые имеют высокую экспрессию на определенных фазах по базе данных Cyclebase. Эти результаты предполагают, что клеточный цикл представляет собой немаловажный источник транскрипционной гетерогенности в культивируемых ADSC, и скрытая гетерогенность может быть незаметна [221].

В последнем были проанализированы транскрипционные состояния клонов, происходящих из нервного гребня и мезодермы, дифференцирующихся в надпочечники, почки, эндотелий и кроветворную ткань между 6 и 14 неделями развития человека после зачатия. Результаты выявляют переходы, связывающие промежуточную мезодерму и предшественников зачатков органов, гематопоэтическую систему и подтипы эндотелиальных клеток. Используя комбинацию транскриптомики единичных клеток и отслеживания клонов, было обнаружено, что внутринадпочечниковые симпатобласты на этой стадии непосредственно происходят из нервно-ассоциированных предшественников шванновских клеток, подобно локальным хромаффинным клеткам, тогда как большинство вненадпочечниковых симпатобластов возникают из мигрирующего нервного гребня. У человека этот процесс сохраняется в течение нескольких недель развития в крупных структурах, подобных ганглиям надпочечников, которые также могут служить резервуарами исходных клеток в нейробластоме [222].

Объединение данных транскриптомики единичных клеток с доступностью хроматина позволяет глубже понять природу гетерогенности

клеток в популяции. Было проведено одноядерное секвенирование ATAC (snATAC-seq) и РНК (snRNA-seq) для создания парных, специфичных для клеточного типа профилей доступности хроматина и транскрипционных профилей почек взрослого человека. Показано, что snATAC-seq сравним с snRNA-seq в определении идентичности клеток и может дополнять понимание функциональной гетерогенности нефрона. Большинство дифференциально доступных областей хроматина локализовано на промоторах, и значительная их часть тесно связана с дифференциально экспрессируемыми генами. Специфичное для клеточного типа обогащение мотивов связывания факторов транскрипции подразумевает активацию NF-κB, которое способствует экспрессии VCAM1 и управляет переходом между субпопуляциями эпителиальных клеток проксимальных канальцев. Многопрофильный подход улучшает способность обнаруживать уникальные состояния клеток в почках и переопределяет клеточную гетерогенность в проксимальных канальцах и толстой части восходящей петли Генле [223].

Идентификация цис-регуляторных элементов, контролирующих паттерны экспрессии специфических генов клеточного типа, важна для понимания происхождения клеточной гетерогенности. Обычным методам картирования регуляторных элементов с помощью анализа открытого хроматина первичных тканей препятствует гетерогенность образца. Одноклеточный анализ доступного хроматина (scATAC-seq) может преодолеть это ограничение. Однако высокий уровень шума профиля каждой отдельной клетки и большой объем данных создают новые ресурсоемкие вычислительные задачи. Для анализа клеточной гетерогенности и построения траекторий клеточных состояний был использован метод SnapATAC. Использование метода Нистрома в SnapATAC может обрабатывать данные из миллиона клеток. Кроме того, SnapATAC объединяет существующие инструменты в комплексный пакет для анализа набора данных ATAC-seq одной клетки. В качестве демонстрации SnapATAC использовался в анализе 55 592 одноядерных профилей ATAC-seq из вторичной моторной коры

головного мозга мыши. Анализ выявил ~ 370 000 кандидатов регуляторных элементов из 31 отдельной клеточной популяции в этой области мозга и, что позволило предположить наличие специфических регуляторов транскрипции для конкретных типов клеток [224].

Анализ графиков RNA velocity дает много информации о транскрипционной динамике клетки. Она позволяет прогнозировать будущее состояние клетки, основываясь на соотношении различных форм РНК, динамике поведения отдельных генов.

Чтобы получить представление о том, как интерпретировать фазовый портрет со сращиванием и без сращивания. Активность генов регулируется транскрипционная регуляция. Индукция транскрипции для определенного гена приводит к увеличению (вновь транскрибируемых) предшественников несплайсированных мРНК, в то время как, наоборот, репрессия или отсутствие транскрипции приводит к снижению несплайсированных мРНК. Сплайсированная мРНК производится из несплайсированной мРНК и следует той же тенденции с задержкой во времени. Время – это скрытая переменная. Таким образом, динамику необходимо выводить из того, что фактически измеряется: сплайсированные и несплайсированные формы мРНК.

Черная линия соответствует расчетному «установившемуся» соотношению, то есть отношению количества несплайсированных и сплайсированных мРНК, который находится в постоянном транскрипционном состоянии. Скорость РНК для конкретного гена определяется как соотношение, т.е. насколько наблюдение отклоняется от этой установившейся линии. Положительная скорость указывает на то, что ген активирован, что происходит в клетках, которые показывают более высокое содержание несплайсированной мРНК для этого гена, чем ожидалось в устойчивом состоянии. И наоборот, отрицательная скорость указывает на то, что ген подавлен.

Глава 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1 Выделение МСК

МСК были выделены из подкожной жировой ткани, полученной из абдоминальной области в ходе хирургической операции у молодого, относительно здорового донора (мужчина, 38 лет).

В дальнейшем анализ был повторен с использованием клеток другого донора (мужчина, 49 лет). В этом эксперименте МСК выделяли из подкожной жировой ткани, полученной из областей колена и живота.

2.2 Культивирование МСК

В качестве объекта исследований использовали МСК человека, выделенные из жировой ткани здоровых доноров. МСК были выделены из подкожной жировой ткани, полученной из абдоминальной области в ходе хирургической операции у молодого, относительно здорового донора (мужчина, 38 лет). Все процедуры, выполненные с образцами тканей пациентов, соответствовали Хельсинкской декларации и были одобрены Комитетом по этике МНОЦ МГУ имени М.В. Ломоносова (IRB00010587), протокол № 4 (2018).

Клетки 1 пассажа высаживали в 90-100% плотности на 6-луночные планшеты для получения образцов контрольных клеток, которые были культивированы в стандартных условиях (контрольный образец), и клеток, культивированных в условиях профибротического микроокружения (на ранее разработанной в лаборатории модели). Непосредственно перед анализом клетки открепляли от культурального пластика, оценивали их концентрацию в суспензии и жизнеспособность с помощью окрашивания трипановым синим на приборе Countess II (Invitrogen).

Клетки 1 пассажа высаживали в 90-100% плотности на 6-луночные планшеты для получения **2 образцов**: клетки, культивируемые в стандартных условиях (контрольный образец), клетки культивируемые в профибротических условиях. Непосредственно перед анализом клетки

открепляли от культурального пластика, оценивали их концентрацию в суспензии и жизнеспособность с помощью окрашивания трипановым синим на приборе Countess II (Invitrogen).

Клетки 1 пассажа высаживали в 90-100% плотности на 6-луночные планшеты для получения **3 образцов**: МСК из живота, культивированные в стандартных условиях (контрольные образцы); МСК из живота, культивированные в профибротических условиях; МСК из живота, культивированные в присутствии 5 нг/мл TGF-beta1. Непосредственно перед анализом клетки открепляли от культурального пластика, оценивали их концентрацию в суспензии и жизнеспособность с помощью окрашивания трипановым синим на приборе Countess II (Invitrogen).

2.3 Профибротическая модель

Для воссоздания действия на МСК профибротического внеклеточного матрикса (ВКМ) были получены клеточные пласты из фибробластов после стимуляции отложения этими клетками коллагена 1 типа по отработанному нами ранее протоколу и проведена децеллюляризация этих конструкций с дальнейшим культивированием на них МСК с добавлением ключевого профибротического фактора – тромбоцитарного фактора роста бета (TGFb). Ранее было показано, что если культивировать фибробласты в таких условиях, то происходит их дифференцировка в миофибробласты, что позволило считать это моделью профибротического окружения. МСК в таких условиях частично приобретают миофибробластный фенотип, а состав их секретома изменяется в сторону профибротического.

Первичные МСК жировой ткани человека культивировали в полной среде роста (AdvanceStem (HyClone) с добавлением 10% ростовых добавок AdvanceStem supplement (HyClone) и 1% пенициллин-стрептомицином (Gibco)) в стандартных условиях (CO2 5%, Tº - 37ºC) в течение 2-4 пассажей. Клетки снимали с пластика последовательной обработкой Версеном (ПанЭко), а затем ферментом QTase (HyClone). Полученную клеточную

суспензию центрифугировали для осаждения клеток. Клеточный осадок ресуспендировали в полной среде роста а затем высаживали из расчёта 60 тыс. кл./мл. После высаживания клеток ждали их полной адгезии в течение 4-6 часов и затем отмывали культуру от среды бессывороточным раствором DMEM LG (Gibco) с добавлением 1% пенициллина-стрептомицина (Gibco) один раз. После этого добавляли среду для кондиционирования DMEM LG (Gibco) с добавлением 1% пенициллина-стрептомицина (Gibco) и оставляли клетки в инкубаторе на 4 суток.

Для формирования профибротического микроокружения клетки высаживали на децеллюляризованный внеклеточный матрикс дермальных фибробластов человека. Децеллюляризацию проводили по отработанному ранее протоколу с использованием детергента CHAPS и обработкой полученных бесклеточных конструкций ДНКазой I типа. К среде для кондиционирования в таких клетках добавляли 5 нг/мл TGFb (CellSignaling). В качестве контроля использовали клетки, посаженные на культуральный пластик (т.н. нормальное микроокружение). Через 4 дня инкубации отбирали из культуры полученную кондиционированную среду. Для удаления клеточного дебриса кондиционированную среду центрифугировали в режиме 300g 10 мин.

2.4 Пробоподготовка

Для дальнейшего анализа использовали от 4 до 10 тысяч живых клеток. С использованием данных клеточных образцов по протоколу производителя коммерческого набора для 10x Genomics нами были подготовлены библиотеки для высокопроизводительного секвенирования.

2.5 Анализ качества библиотек

Содержание кДНК в образцах анализировали с помощью Qubit с использованием набора Qubit DNA HS Assay Kit, для количественного определения ДНК (Thermo Fisher Scientific, Q32851). Качество полученных

библиотек оценивали на приборе Bioanalyzer2100 с использованием набора реагентов High Sensitivity DNA Kit (Agilent Technologies, 5067-4626) (Рис. 23).

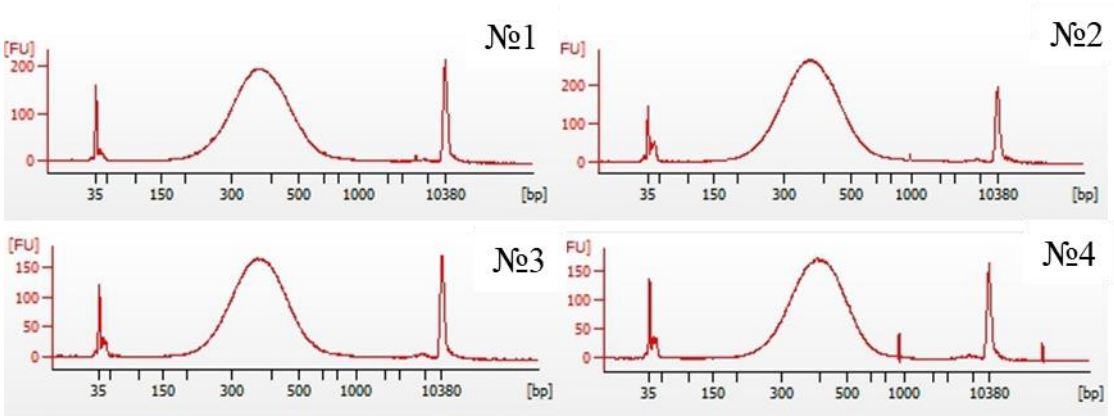


Рисунок 23. Анализ качества библиотек транскриптов.

№ 1 – МСК, культивируемые в стандартных условиях

№3 – МСК, культивируемые в профибротических условиях

Содержание кДНК в образцах анализировали с помощью Qubit с использованием набора Qubit DNA HS Assay Kit, для количественного определения ДНК (Thermo Fisher Scientific, Q32851). Качество полученных библиотек оценивали на приборе Bioanalyzer2100 с использованием набора реагентов High Sensitivity DNA Kit (Agilent Technologies, 5067-4626) (Рис. 24).

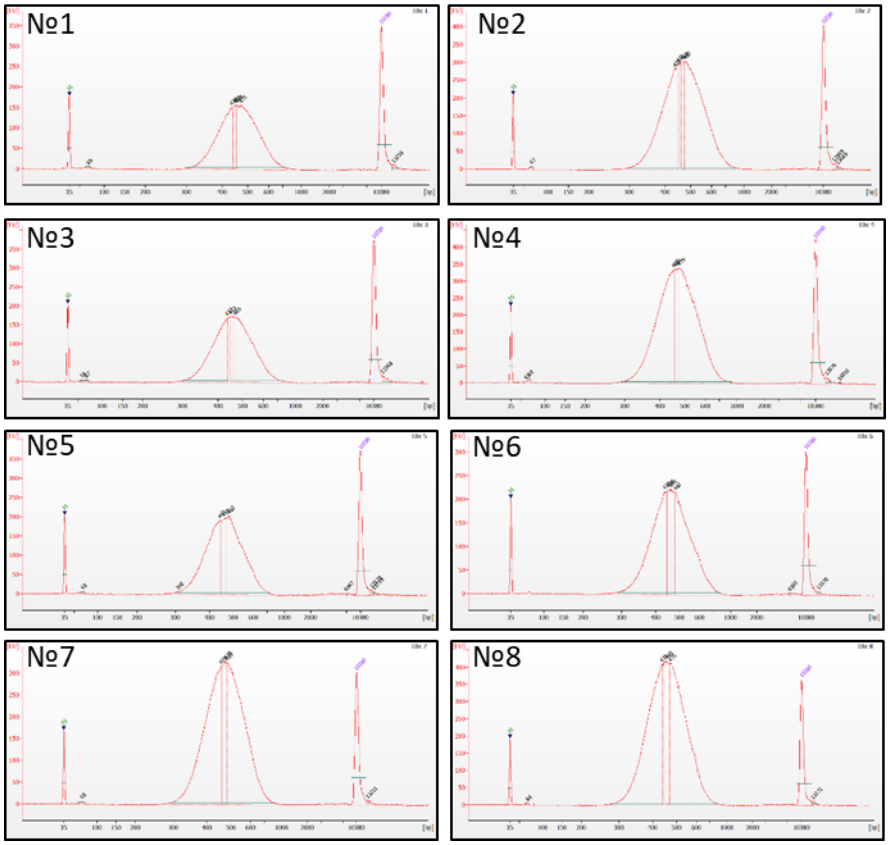


Рисунок 24. Анализ качества библиотек транскриптов.

№ 1 – МСК жировой ткани из абдоминальной области, культивируемые в стандартных условиях

№ 3 – МСК жировой ткани из абдоминальной области, культивируемые в присутствии TGF-beta1

№ 4 – МСК жировой ткани из абдоминальной области, культивируемые в профибротических условиях

Качество всех полученных библиотек соответствовало требованиям, предъявляемым к образцам, направляемым на секвенирование.

2.6 Секвенирование

В первом повторе. Секвенирование парноконцевой библиотеки, подготовленной на приборе Chromium (10x Genomics) было проведено на HiSeq1500 (Illumina), с использованием набора реактивов Chromium Next GEM Single Cell 3’ GEM, Library & Gel Bead Kit (10x Genomics), с длиной прочтений 150 нуклеотидов. Среднее число чтений на образец составило. Глубина прочтений 1,2 млрд. пар чтений (20000 прочтений на клетку). В

результате секвенирования было получено 32 файла, для каждого образца по 8 файлов. Для каждого образца было выполнено 4 технических повтора по два прочтения в каждую сторону. Затем данные секвенирования были подвергнуты оценке контроля качества и биоинформатическому анализу.

Во втором повторе. Секвенирование парноконцевой библиотеки, подготовленной на приборе Chromium (10x Genomics) было проведено на NovaSeq 6000 с использованием реактивов S4, с длиной прочтений 150 нуклеотидов.

Каждая молекула в образце пронумерована UMI, обычно случайной олигонуклеотидной последовательностью, до проведения этапа ПЦР. Когда в процессе анализа обнаруживаются две одинаковые метки на двух идентичных последовательностях, у них одна и та же исходная молекула. Если обнаруживаются две разных метки на одной и той же последовательности, это означает, что это были две разные молекулы.

Большинство экспериментов по секвенированию РНК проводятся на оборудовании, которое предназначено для секвенирования молекул ДНК. В связи с этим необходимым шагом для секвенирования РНК является создание библиотеки кДНК (комплементарной ДНК), полученной из исследуемой тотальной РНК. Каждая кДНК из такой библиотеки представляет собой фрагмент ДНК разного размера, фланкированный по обоим краям специальными адаптерами. Наличие адаптеров необходимо для последующей амплификации образцов и секвенирования. Методы создания библиотек кДНК варьируются в зависимости от конечной цели исследования и типа изучаемой РНК (РНК может различаться в размере, последовательности, структурных особенностях, а также в концентрации). Перед созданием библиотеки кДНК, подходящей для конкретного эксперимента, необходимо ответить на следующие вопросы: 1) какие именно молекулы РНК представляют интерес; 2) как получить кДНК желаемого размера; 3) каким способом лучше присоединить адаптерные последовательности к краям кДНК для амплификации и секвенирования.

Перед секвенированием кДНК ее необходимо амплифицировать с помощью ПЦР. Непосредственно перед проведением ПЦР можно ввести молекулярные маркеры. Эта процедура особенно актуальна, если РНК в образце изначально немного, как, например, в случае секвенирования РНК одной клетки.

Метод секвенирования РНК становится основным методом определения того, какие гены и на каком уровне экспрессируются в клетке. С помощью РНК секвенирования можно определять различия в экспрессии генов на различных стадиях развития организма или в разных тканях.

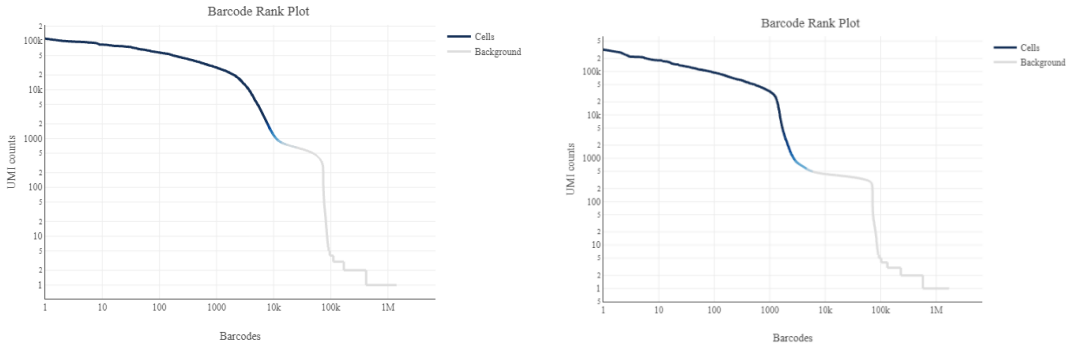
2.7 Биоинформатический анализ данных scRNA-seq

2.7.1 Демультиплексирование и тримминг bcl в fastq

Демультиплексирование файлов bcl полученных после секвенирования проводилось с помощью bcl2fastq, алгоритм которой предполагает удаление адаптерных последовательностей.

2.7.2 Оценка качества прочтений

Для оценки качества использовался файл web-summary из файлов выдачи Cell Ranger (Рис. 25).



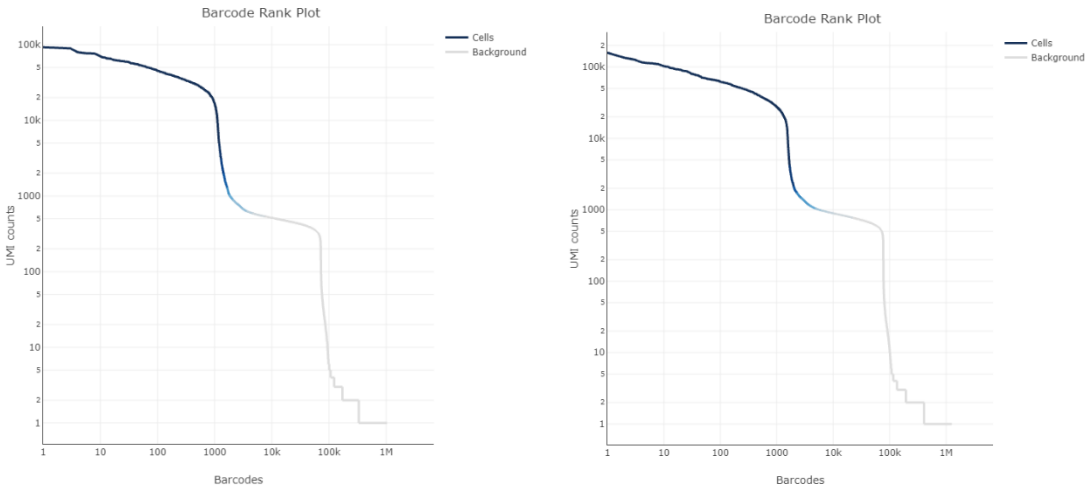


Рисунок 25. Графики Barcode Rank Plot для образцов первого и второго повторов

2.7.3 Выбор протокола анализа данных scRNA-seq

Для анализа fastq-файлов был выбран один из четырех протоколов обработки данных с помощью программного конвейера Cell Ranger «One Sample, One GEM well, Multiple Flowcells», схема которого приведена ниже (Рис. 26).

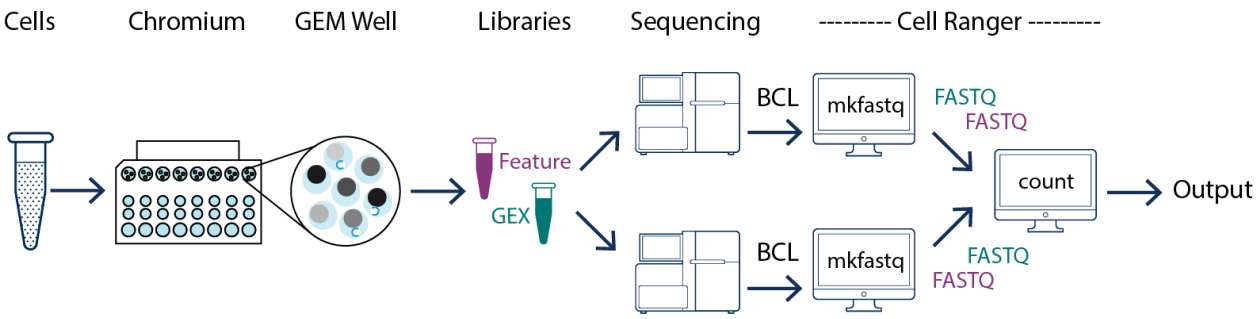


Рисунок 26. Протокол анализа данных scRNA-seq

2.7.4 Картирование на геном и транскриптом

GRCh38 референсный геном и Ensembl 93 референсный транскриптом. Картирование на геномную библиотеку в алгоритме Cell Ranger осуществляется с помощью выравнивателя STAR 2.7.2a. На первом этапе работы программы используются .fastq-файл (библиотека последовательностей) и .gtf-файл (файл с аннотацией) для индексации. На

втором этапе STAR картирует индексированные последовательности на геном генерируя выходные файлы .sam и .bam, содержащие информацию о статистике картирования, выравнивания транскриптов на границы между интронами и экзонами, некартированные риды и т.д.

2.7.5 Картирование на кастомизированный геном

У зрелых микроРНК отсутствует поли-А хвост, поэтому они не могут быть определены в процессе секвенирования. Пре-микроРНК образуются из их предшественников – при-микроРНК. При-микроРНК содержат 5'-кэп и 3' поли-А хвост, поэтому они могут быть транксрибированы в процессе получения библиотеки 10x. В пайплайне Cell Ranger есть возможность добавления необходимых групп генов, например для захвата при-микроРНК или пре-мРНК, в предустановленный референсный геном. Для этого необходимо скачать fasta и GTF файлы соответствующего набора генов и указать пути к этим референсам в командной строке при запуске программы. Также есть возможность указать опцию include-introns, в этом случае дополнения референса не потребуется.

2.7.6 Получение матрицы экспрессии генов

Алгоритм Cell Ranger генерирует два вида матриц в двух форматах. Нефильтрованная матрица содержит все баркоды из фиксированного списка известных последовательностей баркодов, которые имеют хотя бы одно прочтение. Сюда входят и фоновые и ассоциированные с клеткой баркоды. Фильтрованная матрица содержит только баркоды ассоциированные с клеткой.

2.7.7 Нормализация, шкалирование и batch effect

Нормализация проводилась в пакете Seurat с помощью функции NormalizeData. Использовались аргументы normalization.method = "LogNormalize", scale.factor = 10000. Нормализация заключается в делении количества прочтений одного гена для одной клетки на общее количество

прочтений всех генов для данной клетки, умноженное на шкалирующий фактор – 10000. Шкалирование проводилось с помощью функции ScaleData.

2.7.8 Снижение размерности

Снижение размерности в R пакете Seurat производилось с помощью функции RunPCA. Из аргументов функции использовались npcs для указания количества главных компонент для расчета. В большинстве случаев использовалось значение аргумента npcs = 30. Аргумент assay использовался для указания набора генов для отдельного объекта assay = “RNA”, и для интегрированного объекта assay = “Integrated”. Аргумент features для указания набора генов.

2.7.9 Кластеризация

Кластеризация проводилась в R пакете Seurat с помощью функции FindNeighbors. Из аргументов использовался с указанием количества главных компонент dims = 1:10. Эта функция строит граф K-nearest neighbor (KNN) с ребрами, проведенными между клетками с похожими паттернами экспрессии генов, а затем разделяет этот граф на сильно взаимосвязанные "квази-клики" или "сообщества".

Также использовалась функция FindClusters. Из аргументов использовался со значением resolution = 0.5. Для кластеризации клеток применяются методы оптимизации модульности, такие как Лувенский алгоритм (по умолчанию) или SLM, чтобы итеративно сгруппировать клетки вместе, с целью оптимизации стандартной функции модульности.

2.7.10 Дифференциальная экспрессия

Для получения списка дифференциально экспрессирующихся генов в отдельном кластере использовалась функция FindMarkers. Для этой функции использовался аргумент slot для указания раздела объекта, где содержатся значения.

2.7.11 Интеграция scRNA-seq датасетов

Интеграция выполнялась в два этапа: с помощью функции FindIntegrationAnchors определение «anchors» (якорей – пар похожих клеток) и с помощью функции IntegrateData интеграция датасетов.

2.7.12 Типирование клеток

2.7.12.1 Автоматическое типирование клеток

Для автоматического типирования клеток использовался R-пакет SingleR, использующий пакет celldex для доступа к клеточным референсам Human Primary Cell Atlas (HPCA) и Blueprint.

Клеточный референс BlueprintEncodeData содержит нормализованные данные экспрессии 259 образцов bulk RNA-seq из популяций стромальных и иммунных клеток.

Клеточный референс HumanPrimaryCellAtlasData содержит нормализованные данные 713 микрочиповых образцов. Образцы приведены к 37 основным клеточным типам и 157 подтипам.

2.7.12.2 Типирование клеток по специфичным маркерам

Для типирования были использованы функции R-пакета Seurat FindMarkers (используется для поиска дифференциально экспрессирующихся генов между двумя определенными группами клеток – кластерами) и FindAllMarkers (используется для поиска дифференциально экспрессирующихся генов между анализируемой группой клеток и всеми остальными. Затем таблица с дифференциально экспрессируемыми генами была экспортирована в excel с помощью функции write.xlsx. После этого был проведен поиск топовых генов по базам данных маркеров клеточных типов (PanglaoDB [225], CellMarker [226]).

2.7.12.3 Типирование промежуточных форм клеток по биологическим процессам

В некоторых кластерах среди дифференциально экспрессирующихся генов не оказалось специфических маркеров. Получить информацию о процессах происходящих в кластерах с не идентифицированным типом клеток можно по GO: Biological process с помощью он-лайн сервиса String (version 11.5) [227].

2.7.13 Траектории развития

Траектории развития были получены с помощью пакета Dynverse. Из пакета использовались библиотеки dynwrap, dyno, dyndimred, dynplot, dymRed. С помощью функции wrap_expression был создан dyno-объект, содержащий данные о прочтениях и уровнях экспрессии. Затем с помощью функции add_prior_information к объекту была добавлена информация о начальных и конечных точках, add_dimred – информация о координатах клеток, add_grouping – информация о кластерах клеток. С помощью функции infer_trajectory были получены траектории в моделях slingshot и paga-tree. Визуализация траекторий осуществлялась с помощью функции plot_dimred.

2.7.14 RNA-velocity

Матрицы со сплайсированными и несплайсированными прочтениями (loom-файлы) были получены с помощью velocyto, координаты клеток были экспортированы из Seurat, объединение информации было выполнено с помощью scVelo.

2.7.15 In silico поиск микроРНК, вовлечённых в ответ субпопуляций МСК на профибротическое окружение

Дальнейший анализ был проведён с помощью кода, написанного на языке Python. Методами pandas были загружены таблицы, полученные в результате семантического анализа. Было осуществлён отбор дифференциально экспрессированных генов, принадлежащих категориям в

терминах GO, Reactome и KEGG, содержащим следующие ключевые слова: 'cellular', 'stem', 'fibrosis', 'wound', 'differentitation', 'actin', 'matrix', 'fibroblast', 'myofibroblast', 'myofibril', 'wounding', 'muscle', 'extracellular', 'encapsulating', 'mesenchymal', 'collagen', 'contraction', 'ECM', 'Matrix Metalloproteinases'.

Далее методами pandas была загружена база данных mirnet в виде таблицы csv. Был осуществлён поиск микроРНК, для которых отобранные гены являлись мишенями, и создано два списка, в один из которых входили микроРНК, имеющие в качестве мишеней гены с повышенной экспрессией, в другой - гены с пониженной экспрессией. Затем по полученным спискам была произведена обратная процедура поиска всех возможных генов-мишеней данных микроРНК. В отдельную категорию были выделены микроРНК, имеющие мишени как среди генов с повышенной экспрессией, так и среди генов с пониженной экспрессией. Будем называть их микроРНК с перекрывающейся специфичностью (overlapped). Было посчитано общее число мишеней для каждой микроРНК, а также специфичность данных микроРНК по отношению к отобранным генам как процент количества генов из числа отобранных от общего числа генов-мишеней данных микроРНК. Полученные результаты были записаны в единую таблицу формата .xlsx.

Следующим этапом стал отбор микроРНК в соответствии с результатами таргетного ПЦР. Отбирались микроРНК, для которых был проведён данный анализ. Затем гены-мишени данных микроРНК были проаннотированы по базам данных KEGG и GO с помощью функции GProfiler.profile с параметрами: организм, для генов которого осуществляется аннотация (organism) hsapiens, что значит Homo sapiens, человек разумный, выбор баз данных (sources) KEGG, GO. Далее полученные термины были отфильтрованы по наличию в них ключевых слов: 'stem', 'fibrosis', 'wound', 'differentitation', 'actin', 'matrix', 'fibroblast', 'myofibroblast', 'aging', 'senescence', 'apoptosis', 'wounding', 'muscle', 'programmed', 'extracellular', 'encapsulating', 'proliferation', 'mesenchymal'. Полученный результат был записан в таблицу формата .xlsx. Для отобранных микроРНК был проведён анализ литературы

на предмет связи данных микроРНК с фибротическими состояниями и их регуляторным воздействием на МСК.

2.7.16 Поиск антифибротических регулонов

Пакет Scenic позволяет выявить коэкспрессирующиеся транскрипционные факторы для анализа их участия в регуляции процессов в клетке. Алгоритм включает в себя выполнение команд трех пакетов. GENIE3 для идентификации транскрипционных коэкспрессирующихся факторов, RcisTarget для определения прямых мишеней (регулонов) транскрипционных факторов и анализа обогащения мотивами транскрипционных факторов, AUCell для того, чтобы рассчитать активность регулонов относительно единичных клеток [228].

Для получения списка регулонов был использован пакет Scenic 1.1.2.

Глава 3. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

3.1 Клеточная модель

Для индукции дифференцировки фибробластов в миофибробласты клетки высаживали в культуральные планшеты из расчёта 15 тыс/см2 и культивировали в течение суток. Депривировали культуру в бессывороточной среде ДМЕМ (Gibco) с содержанием 1% HyClone Penicillin-Streptomycin 100X Solution (HyClone) в течение ночи. После депривации меняли в клетках среду на фракцию, обогащённую ВВ-МСК (1,75 × 105 част/кл.) или РФ-МСК (сконцентрированную до аналогичного ВВ-МСК объёма). Для индукции дифференцировки в миофибробласты одновременно с ВВ-МСК или РФ-МСК в среду вносили 5 нг/мл TGFβ (R&D, США). ДМЕМ без ФБС и без TGFβ использовали в качестве отрицательного контроля (группа «контроль»); ДМЕМ без ФБС с 5 нг/мл TGFβ использовали в качестве положительного контроля (группа «+TGFβ»). Клетки помещали в CO2-инкубатор при 37 °C и анализировали через 4 дня.

Для создания модели дедифференцировки миофибробластов в лунки планшета высаживали фибробласты из расчёта 15 тыс/см2. Через сутки культивирования клетки депривировали в бессывороточной среде ДМЕМ (Gibco) с содержанием 1% HyClone Penicillin-Streptomycin 100X Solution (HyClone) в течение ночи. После депривации меняли в клетках среду на ДМЕМ с добавлением 5 нг/мл TGFβ. После 4 суток инкубации в культуре, когда клетки приобретали фенотип миофибробластов, меняли среду на фракцию, обогащённую ВВ-МСК или РФ-МСК. ДМЕМ без ФБС и без TGFβ использовали в качестве отрицательного контроля (группа «контроль»). Клетки помещали в CO2-инкубатор при 37 °C и анализировали через 4 дня.

3.2 Гетерогенность

3.2.1 Выявление субпопуляции клеток-нереспондеров в образце МСК, культивируемых под воздействием профибротических стимулов

В результате кластеризации клеток образца МСК, культивируемых в профибротических условиях получено 6 кластеров (Рис. 27).

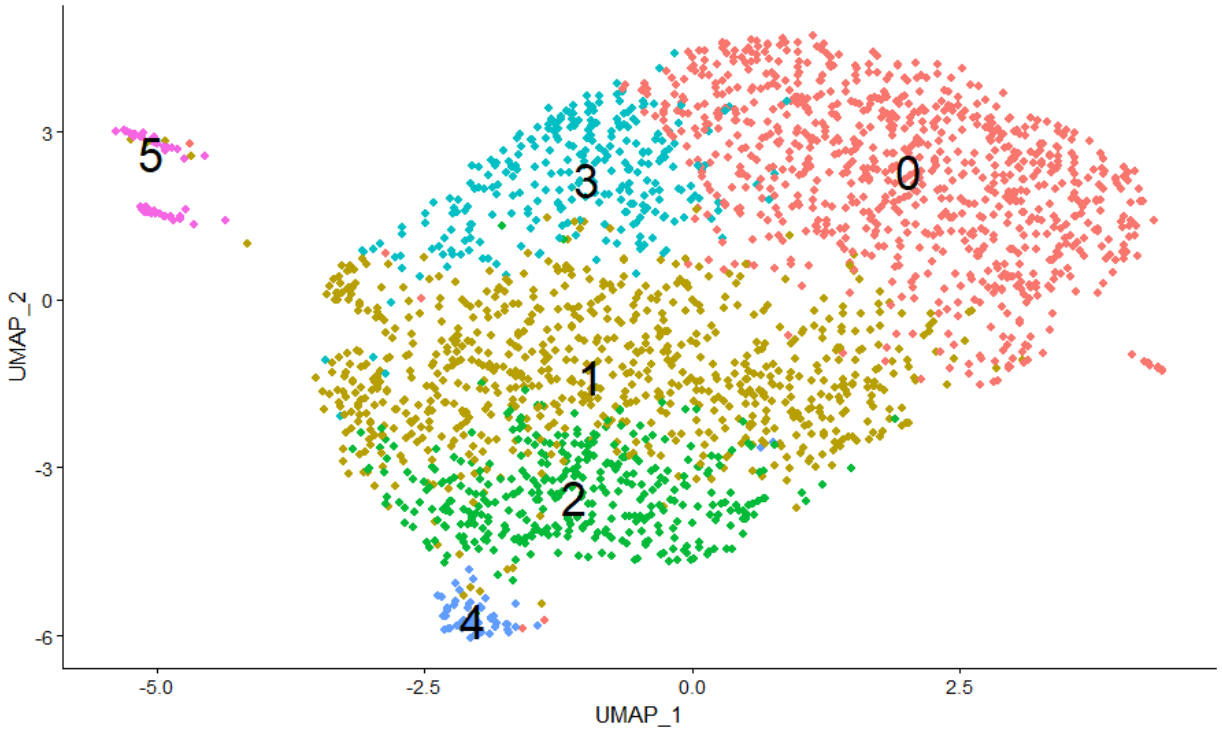


Рисунок 27. Результат кластеризации образца МСК, культивируемых в профибротических условиях (UMAP), N=2.

Для разделения полученных кластеров на про- и антифибротические, в качестве маркера миофибробластной дифференцировки был выбран α-гладкомышечный актин (α-SMA) (Рис. 28).

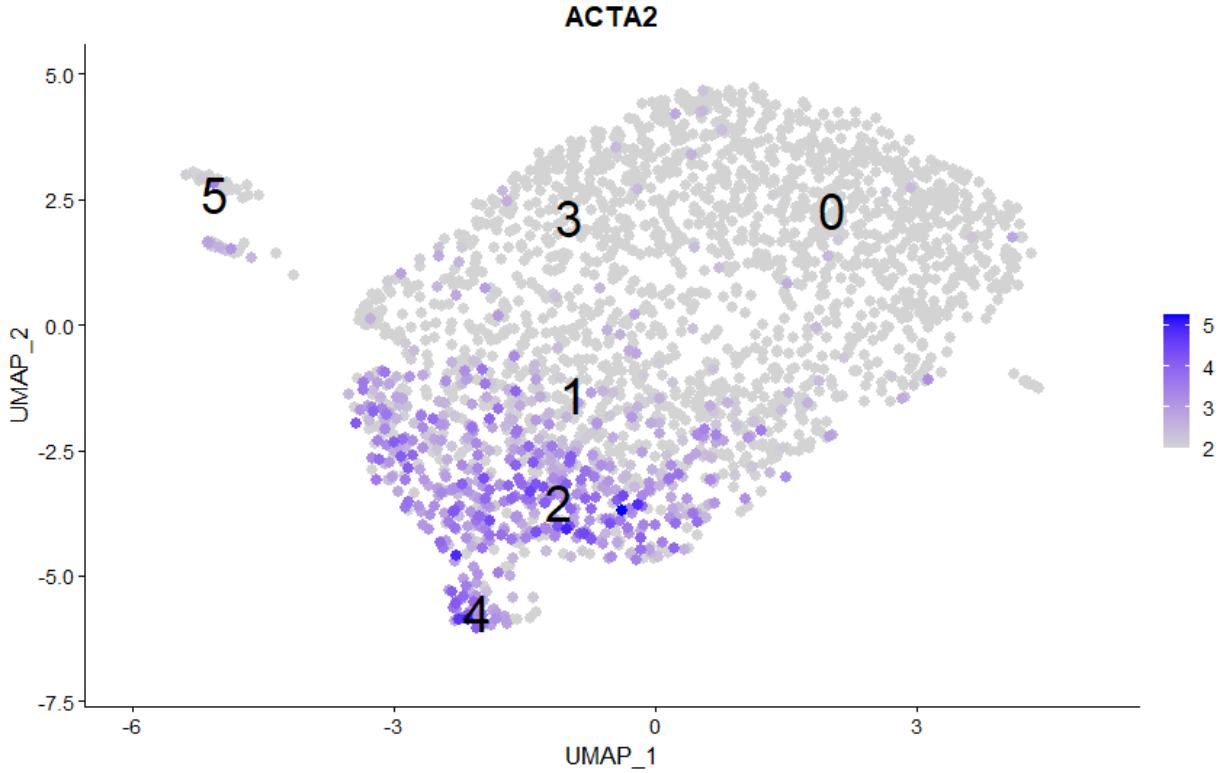


Рисунок 28. Клетки, в которых уровень экспрессии гена ACTA2 (α-SMA) в 5 раз больше по сравнению со всеми МСК образца. N=2.

На графике МСК с высоким уровнем экспрессии гена ACTA2 преимущественно располагаются во 2 и 4 кластерах. Эти кластеры в дальнейшем анализе именуются профибротическими. В 0, 1, 3 и 5 кластерах сосредоточились МСК с низким уровнем экспрессии гена ACTA2. Эти кластеры в дальнейшем анализе именуются антифибротическими.

Для минимизации ошибки определения субпопуляций образца МСК как про- и антифибротические использовалась числовая оценка уровня экспрессии гена ACTA2 в кластерах. В экспортированной из Seurat Object таблице (Табл. 6) проведено сравнение разницы между средними значениями (Average), кратностями изменения (Log_2 Fold Change) уровня экспрессии гена ACTA2 в кластерах образца.

Таблица 6. Кратность изменения уровня экспрессии гена ACTA2 в кластерах образца МСК, культивируемых в профиброгенных условиях относительно всех клеток образца.

Gene	Cluster	avg_log2FC
ACTA2	0	-0.5966739
	1	0.1080111
	2	0.7875731
	3	-0.3638739
	4	0.7385978
	5	-0.2343277

Среднее значение и кратность изменения уровня экспрессии гена ACTA2 в кластерах 2 и 4 в среднем выше более, чем в 7 раз. Следовательно, в дальнейшем анализе кластеры 0, 1, 3 и 5 рассматриваются как антифибротические, а кластеры 2 и 4 – профибротические.

3.2.2 Дифференциальная экспрессия генов функциональных групп образца МСК, культивируемых в профибротических условиях

Известно, что МСК секретируют факторы, которые по механизму действия можно поделить на несколько кластеров: ангиогенные (восстановление трофики ткани), нейротрофные (интеграция восстановленной ткани в регуляторную сеть), иммуномодулирующие (регуляция воспаления и иммунного ответа), матриксные белки (формирующие структуру ткани), а также факторы, стимулирующие метаболизм клетки [229]. В начале рассмотрим профибротические кластеры 2 и 4.

Для функционального анализа 2 кластера были использованы гены с avg_log2FC>0.6 (31 ген) (Табл. 7).

Таблица 7. Функциональное обогащение белков 2 кластера

GO-term	description	count in network	strength	false discovery rate
GO:0006936	Muscle contraction	7/248	1.31	7.16e-05
GO:0009611	Response to wounding	6/532	0.98	0.0397

Примечание: **GO-term** (генная онтология), **description** (название термина), **count in network** (первая цифра – количество генов в анализируемой сети, совпавших со списком всего термина),

strength (Log10(наблюдаемый/ожидаемый). Мера величины эффекта обогащения. Это соотношение между количеством белков в вашей сети, аннотированных термином, и количеством белков, которые, могли бы быть аннотированы этим термином в случайной сети того же размера. **false discovery rate** (значительность обогащения, p-значения, скорректированные на множественное тестирование в каждой категории с помощью процедуры Бенджамина-Хохберга)

Для функционального анализа 4 кластера были использованы гены с avg_log2FC>0.6 (31 ген) (Табл. 8).

Таблица 8. Функциональное обогащение белков 4 кластера

GO-term	description	count in network	strength	false discovery rate
GO:0061041	Regulation of wound healing	4/140	1.27	0.0454
GO:0030334	Regulation of cell migration	6/865	0.68	0.0454

Во 2 и 4 кластерах заметно выражен ответ на повреждение и процессы регуляции заживления ран и клеточной миграции, что является основными процессами при развитии фиброза. Также обнаруживаются гены участвующие в мышечном сокращении.

Для функционального анализа 0 кластера были использованы гены с avg_log2FC>0.4 (31 ген). Для данного кластера был определен только один процесс - GO:0030198 Extracellular matrix organization, 3.87e-05. 8 генов из 338 генов этого класса GO - DCN, PDGFRA, CTSK, COL6A3, MMP3, FBLN1, CTSL, DPT.

Для функционального анализа 1 кластера было использовано 52 гена с avg_log2FC>0.25 (Табл. 9).

Таблица 9. Наиболее представленные процессы 1 кластера в терминах GO

GO-term	description	count in network	strength	false discovery rate
GO:0030198	Extracellular matrix organization	22/338	1.39	7.00E-21
GO:0072359	Circulatory system development	18/872	0.89	1.43E-08
GO:0001568	Blood vessel development	13/500	0.99	9.23E-07
GO:0001501	Skeletal system development	12/499	0.96	5.88E-06
GO:0009888	Tissue development	20/1760	0.63	7.57E-06
GO:0001503	Ossification	9/265	1.11	2.21E-05

GO:0048514	Blood vessel morphogenesis	10/410	0.96	6.13E-05
GO:0061448	Connective tissue development	8/223	1.13	7.24E-05

Важными процессами в 1 кластере являются организация клеточного матрикса, развитие сосудов и системы кровообращения, оссификации, развития опорно-двигательной системы, а также развитие соединительной ткани, что подтверждает имеющиеся знания о направлениях дифференцировки МСК в фибробласты, гладкомышечные клетки, адипоциты, хондробласты и остеобласты. Наличие терминов, связанных с ростом сосудов объясняется участием секретома МСК в этом процессе.

Для функционального анализа 3 кластера было использовано 24 гена, 5 кластера – 20 (Табл. 10).

Таблица 10. Наиболее представленные процессы 3 кластера в терминах GO

GO-term	description	count in network	strength	false discovery rate
GO:0006695	Cholesterol biosynthetic process	6/41	2.08	4.67e-08
GO:0046890	Regulation of lipid biosynthetic process	7/197	1.46	1.75e-06
GO:0062012	Regulation of small molecule metabolic process	8/449	1.16	6.41e-06
GO:0050810	Regulation of steroid biosynthetic process	5/89	1.66	2.70e-05
GO:0045338	Farnesyl diphosphate metabolic process	3/5	2.69	4.27e-05
GO:0045540	Regulation of cholesterol biosynthetic process	4/43	1.88	9.41e-05

В клетках 3 кластера протекают процессы, связанные с обменом липидов, а именно, регуляция биосинтеза холестерина, а также одного из промежуточных продуктов метаболизма холестерина – фарнезилдифосфата.

Для функционального анализа 5 кластера было использовано 20 генов. В 5 кластере активно протекает митоз GO:0051301 Cell division (12/491) 1.02e-10 - CCNB1, TPX2, BIRC5, TUBA1B, TUBB, UBE2C, SMC4, CENPF, CDC20, PTTG1, PRC1, STMN1.

3.2.3 Гетерогенность ответа МСК на TGFb

Фибробласты под воздействием TGFβ и других провоспалительных факторов способны трансдифференцироваться в миофибробласты. Повышенная экспрессия TGFβ стимулирует миофибробласты к гиперсекреции компонентов ВКМ, ингибирует синтез ММР и усиливает производство тканевых ингибиторов металлопротеиназ (TIMPs), увеличивает образование клеточных контактов различного типа, что может приводить к развитию патологического фиброза.

Ниже представлены результаты кластеризации МСК, культивируемых под воздействием TGFb (Рис. 29).

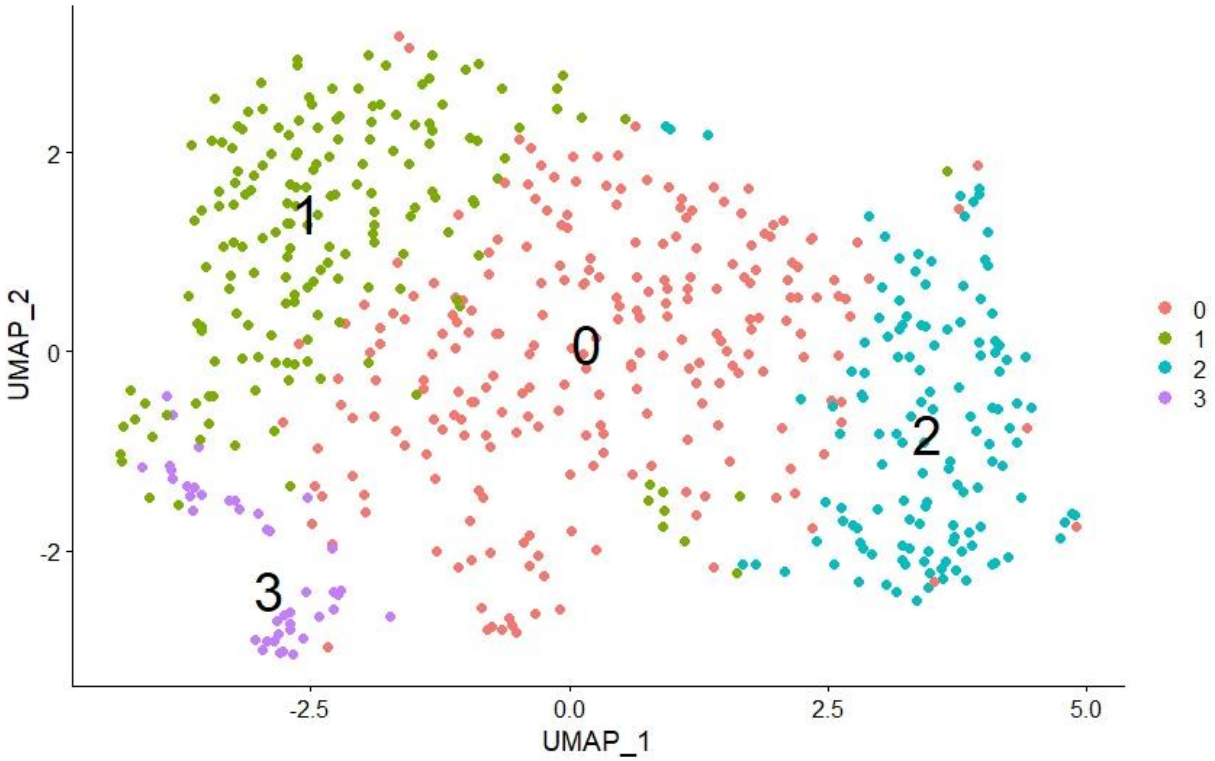


Рисунок 29. Результат кластеризации образца МСК, обработанных TGFb

В образце было 2300 клеток и при кластеризации было получено 4 кластера. Для разделения полученных кластеров на про- и антифибротические, в качестве маркера миофибробластной дифференцировки был выбран α-гладкомышечный актин (α-SMA) (Рис. 30).

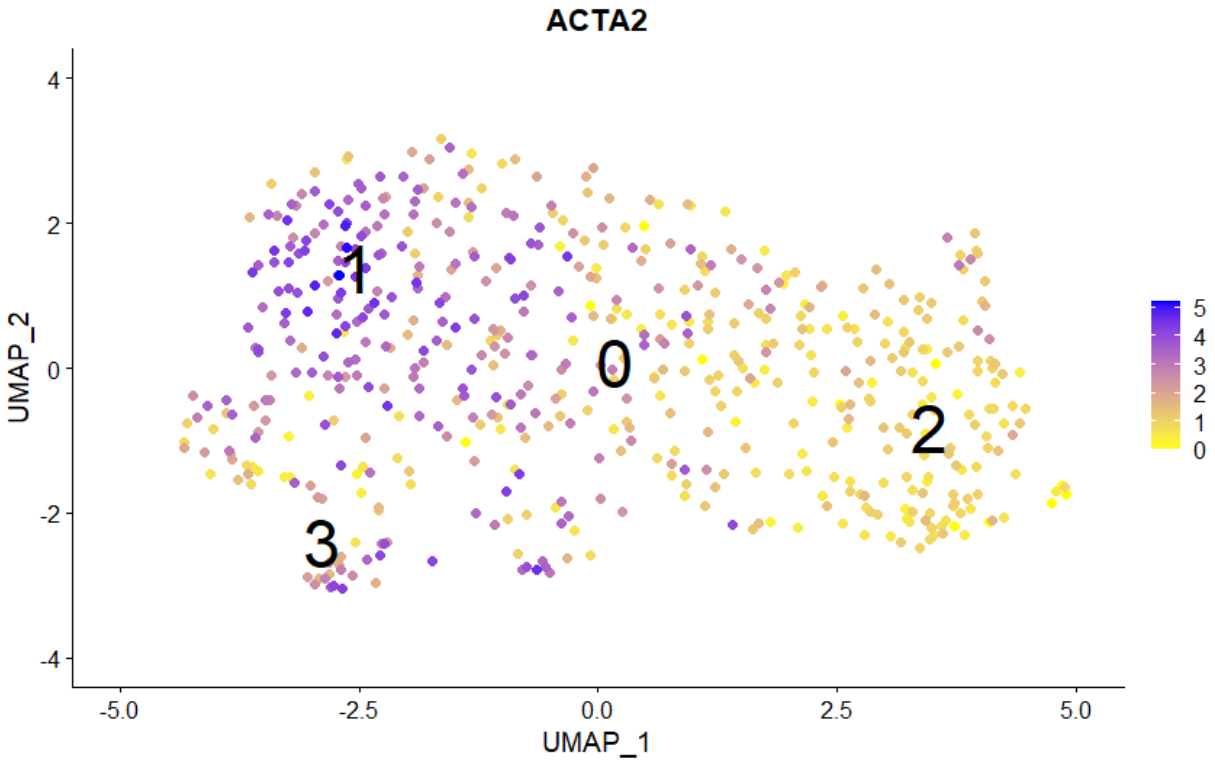


Рисунок 30. Уровень экспрессии (log2) гена ACTA2 (α-SMA) в образце МСК обработанного TGFβ

На рисунке видно, что в образце выделяются две субпопуляции по ключевому маркеру миофибробластов – aSMA. На графике МСК с высоким уровнем экспрессии гена ACTA2 преимущественно располагаются в 1 кластере. В 0, 2 и 3 кластерах сосредоточились МСК с низким уровнем экспрессии гена ACTA2 (Табл. 11).

Таблица 11. Кратность изменения уровня экспрессии гена ACTA2 в кластерах образца МСК, культивируемых под действием TGFb относительно всех клеток образца.

	Cluster	avg_lof2FC
ACTA2	0	-0.6170977
	1	1.6707
	2	-2.845905
	3	0.1501197

Для минимизации ошибки определения субпопуляций образца МСК как про- и антифибротические использовалась числовая оценка уровня экспрессии гена ACTA2 в кластерах. В экспортированной из Seurat Object таблице проведено сравнение разницы между средними значениями

(Average), кратностями изменения (Log2 Fold Change) уровня экспрессии гена ACTA2 в кластерах образца.

3.2.4 Распределение маркеров миофибробластов образца МСК, культивируемых в профибротических условиях

Анализ результатов кластеризации данных scRNA-seq МСК, культивируемых в профибротических условиях и сравнение выделенных субпопуляций между собой проводилось с помощью R-пакета Seurat.

В клеточных референсах PanglaoDB и CellMarker существует информация о маркерах миофибробластах. Для проверки наличия дифференцированных миофибробластов было изучено распределение известных маркеров CALD1, MYL9 и TAGLN [230, 231] (Рис. 31).

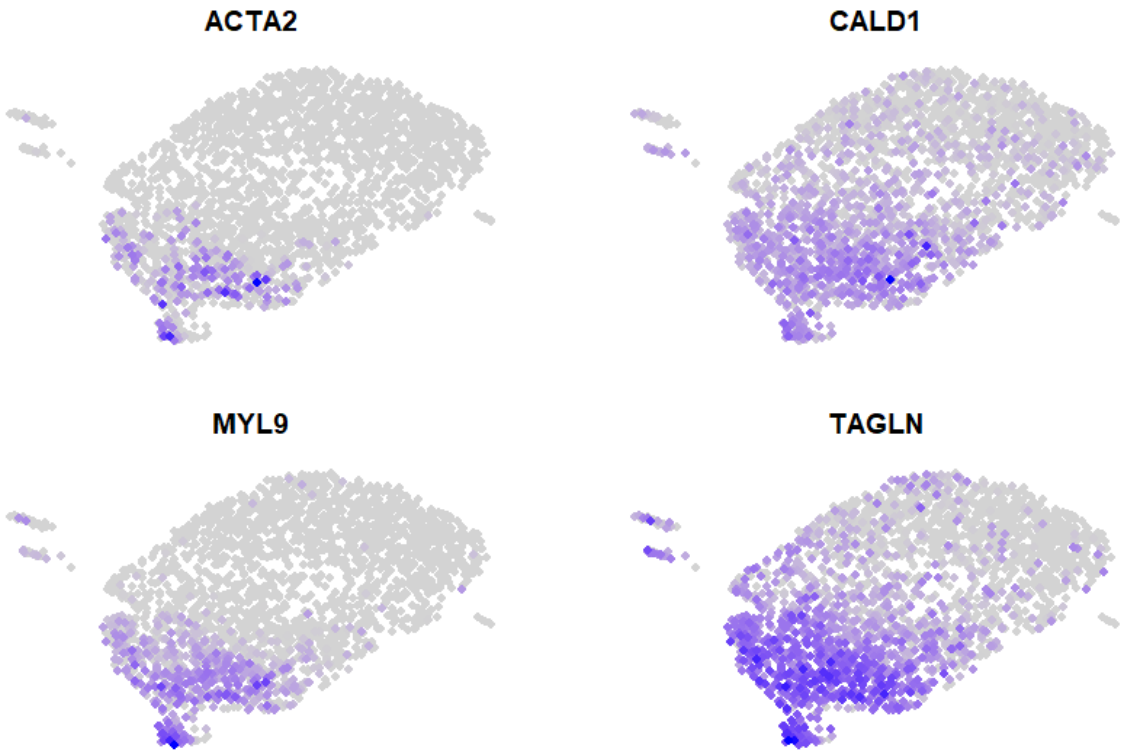


Рисунок 31. Экспрессия маркеров миофибробластов в образце МСК, культивируемых под действием профибротических стимулов

Перечисленные маркеры миофибробластов преимущественно сосредоточены во 2 и 4 кластерах массива, что позволяет предположить, что та часть МСК, которая в течение 4 дней вступила в дифференцировку, располагается именно в них.

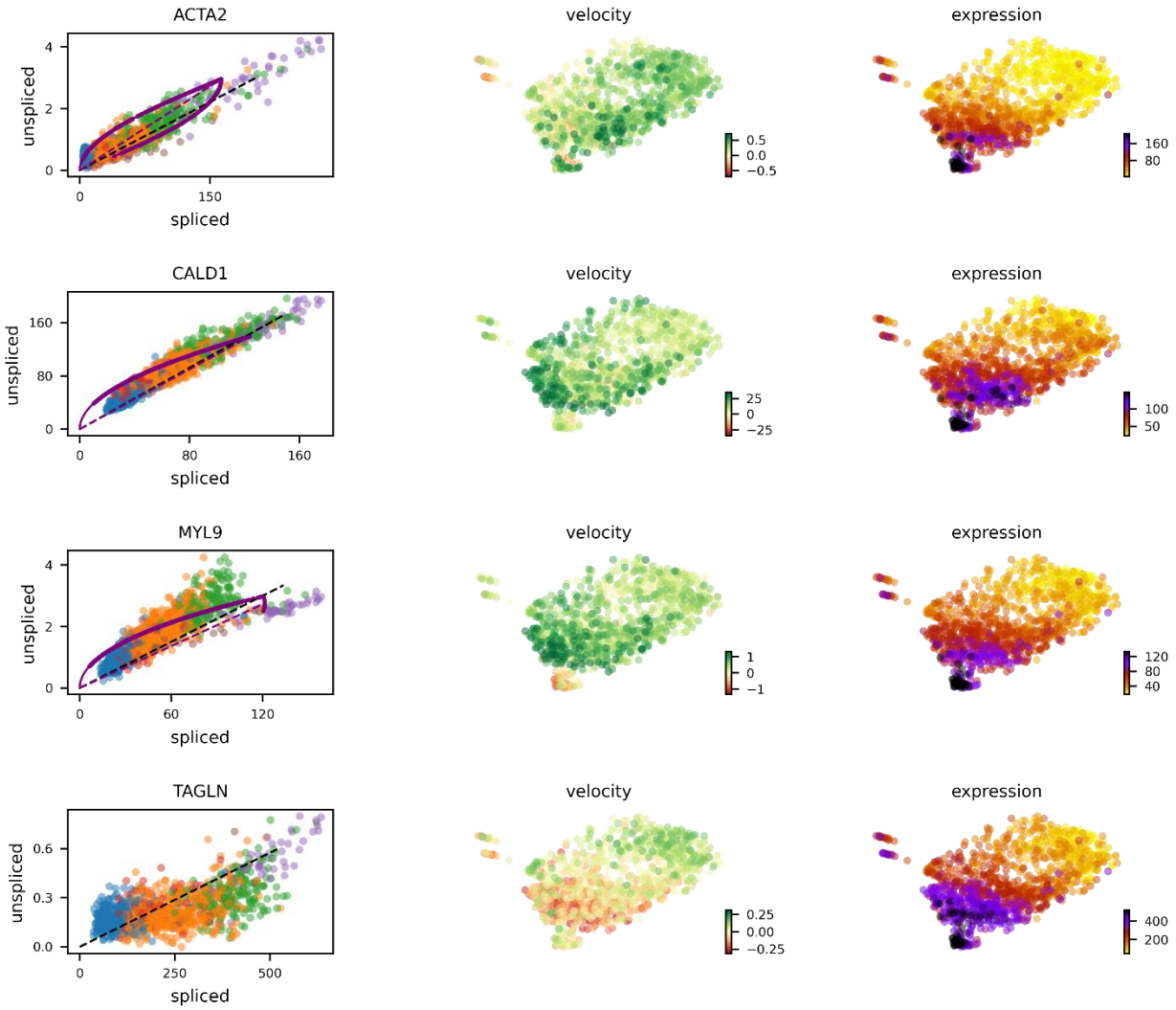


Рисунок 32. Графики соотношения сплайсированных и несплайсированных форм, RNA-velocity и дифференциальной экспрессии маркеров миофибробластов образца МСК, культивируемых под воздействием профибротических стимулов

Анализ соотношения несплайсированных форм транскриптов к сплайсированным, или RNA-velocity, показывает, что экспрессия генов ACTA2, CALD1 и MYL9 активно проходит во 2 и 4 кластерах, а TAGLN начал подвергаться деградации (Рис. 32).

3.2.5 Распределение дифференциально экспрессирующихся генов α -sma$^-$ - субпопуляции

Для того, чтобы охарактеризовать выделенные субпопуляции в них были определены дифференциально экспрессирующиеся гены. Список

высокоэкспрессированных генов α-SMA$^-$ субпопуляции представлен в таблице 12.

Таблица 12. Сравнение кратности изменения дифференциально представленных генов в α-sma$^-$ субпопуляции со всеми клетками образца (0, 1, 3 и 5 кластеры).

Ген	avg_log2FC	кластер	Ген	avg_log2FC	кластер
CTSK	1.550	0	PDGFRA	0.445	0
CTSL	0.551/0.276	0/5	SOX4	0.427	0
MMP1	0.269	0	SPON2	0.397	0
MMP3	0.827	0	PLPP3	0.393	0
FBLN1	0.627	0	DPP4	0.387	0
MARCKSL1	0.618	0	HCFC1R1	0.372	0
LINC01705	0.585	0	TSHZ2	0.343	0
CHI3L1	0.541	0	COLEC12	0.316	0
GPNMB	0.539	0	PLTP	0.291	0
VIM	0.470/0.269	0/3	CLEC2B	0.357	0

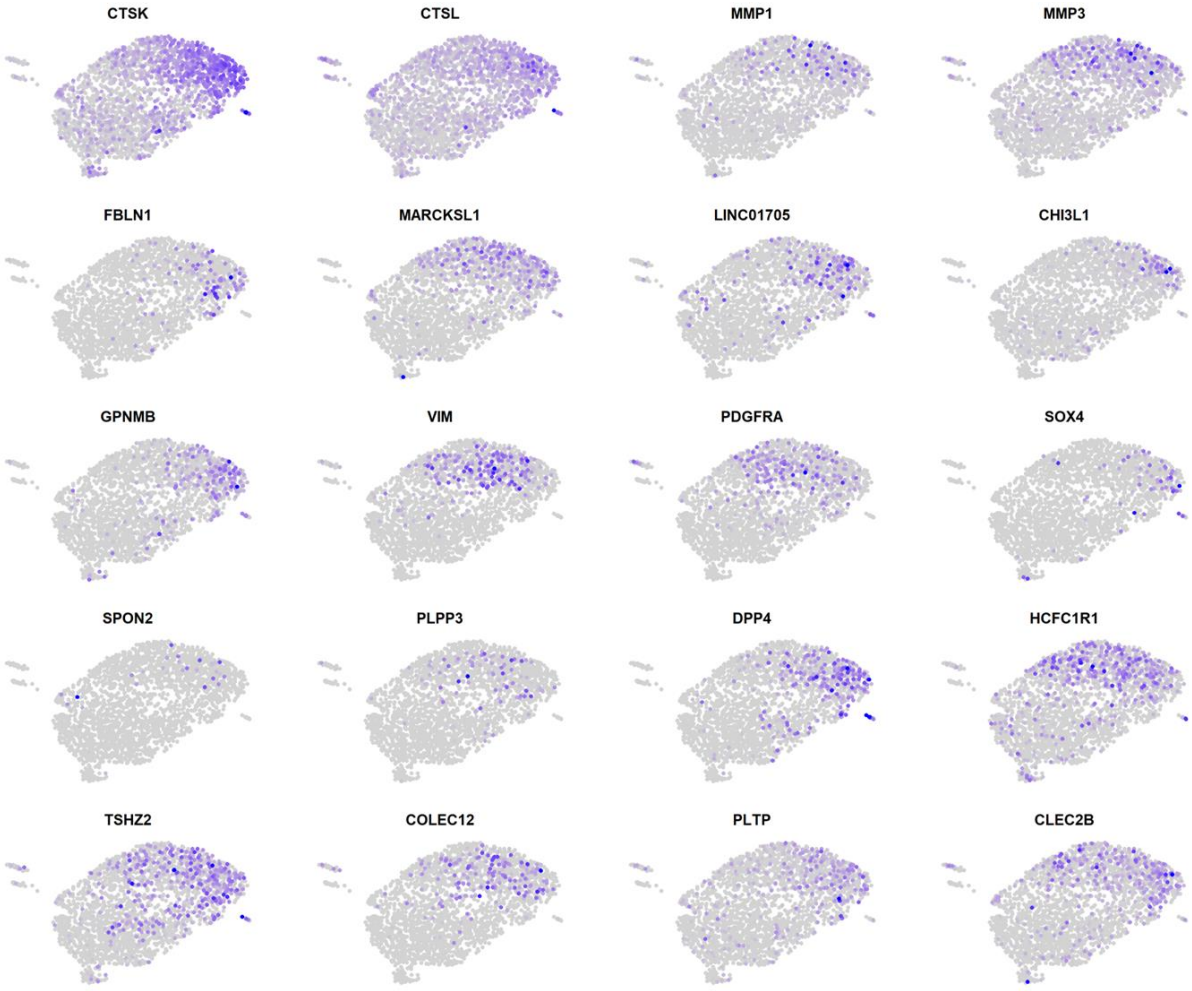


Рисунок 33. Дифференциально экспрессирующиеся гены в α-sma$^-$ - субпопуляции

Катепсин К (CTSK) участвует в разрушении компонентов внеклеточного матрикса [232]. Матриксная металлопептидаза или коллагеназа MMP1 обладает антифибротическим действием благодаря способности расщеплять коллагены 1-3 типов [233]. MMP3 играет важную роль в ремоделировании соединительной ткани, она способна разрушать коллаген (II, III и IV), эластин, фибронектин, ламинин и протеогликан. Кроме того, MMP-3 вызывает активацию MMP-1, MMP-7 и MMP-9 в печени [234]. FBLN1 [235] играет роль в клеточной адгезии и транспорте по волокнам внеклеточного матрикса. MARCKSL1 относится к семейству MARCKS и участвует в реорганизации актинового цитоскелета и миграции фибробластов [236]. Длинная некодирующая РНК LINC01705 взаимодействует с микроРНК, участвующими в регуляции процессов фиброза. С подробным анализом данной днРНК можно ознакомиться в разделе, посвященном участию нкРНК в фиброзе. Роль CHI31L в регуляции фиброза определена не окончательно, однако показано, что ее уровень повышается у пациентов с идиопатическим фиброзом легкого, что может говорить об ответной реакции на повреждение [237]. Выделяемый макрофагами трансмембранный гликопротеин GPNMB задерживается фиброзным внеклеточным матриксом во время транспортировки и может активировать фибробласты через CD44/Serpinb2 путь, что приводит к дальнейшему развитию фиброза [238]. Виментин через различные сигнальные пути контролирует пролиферацию, дифференцировку и движение звездчатых клеток печени через каскады ERK/AKT и Rho участвует в регуляции процессов фиброза [239]. PDGFRA является сигнальной молекулой в регуляции процесса фиброза [240]. Экспрессия и активность транскрипционного фактора развития SOX4 быстро индуцируются на ранних стадиях TGF-β-индуцированного перехода эпителия в мезенхиму (Рис. 33) [241].

Таблица 13. Кластеризация по GO_BP (Gene ontology, biological process) высокоэкспрессированных генов (> 2 раз) в α-sma$^-$ - субпопуляции.

GO-term	description	count in	Padj

		network	
GO:0030154	Cell differentation	32/4440	$2.927*10^{-2}$
GO:0045597	Positive regulation of cell diferentation	13/894	$3.008*10^{-2}$
GO:0097435	Supramolecular fiber organization	12/739	$2.261*10^{-2}$
GO:0030198	Extracellular matrix organization	9/399	$2.522*10^{-2}$
GO:0043062	Extracellular structure organization	9/400	$2.572*10^{-2}$

Высокоэкспрессирующиеся гены входят в состав кластеров GO_BP, объединяющих гены, отвечающие за организации внеклеточного матрикса, что соответствует ожидаемому нами эффекту в α-sma$^-$ субпопуляции – сдерживанию профибротических изменений путем разрушения, образующихся в процессе фиброза внеклеточных матриксных белков (Табл. 13).

3.2.6 Распределение дифференциально экспрессирующихся генов α -sma$^+$ - субпопуляции

Для того, чтобы охарактеризовать выделенные субпопуляции в них были определены дифференциально экспрессирующиеся гены. Список высокоэкспрессированных генов α-SMA$^+$ субпопуляции представлен в таблице 14.

Таблица 14**.** Сравнение кратности изменения дифференциально представленных генов в α-sma$^+$ субпопуляции со всеми клетками образца (2 и 4 кластеры)

Ген	avg_log2FC	кластер	Ген	avg_log2FC	кластер
IGFBP3	1.953	2	CSRP1	1.215	4
CNN1	1.145	2	TM4SF1	0.942	4
SERPINE1	0.960	2	SYNPO2	0.902	4
MYLK	0.947/1.470	2/4	EDIL3	0.899	4
KRT7	0.911	2	LMOD1	0.860	4
CLIC3	0.909	2	CD36	0.828	4
PPME1	0.743	2	TPM2	1.062/ 0.823	2/4
ACTG2	2.103	4	MYH11	0.816	4
APOE	1.921	4	DSTN	0.792	2
COL4A1	1.339	4	CRYAB	0.779	4
PPP1R14A	1.218	4			

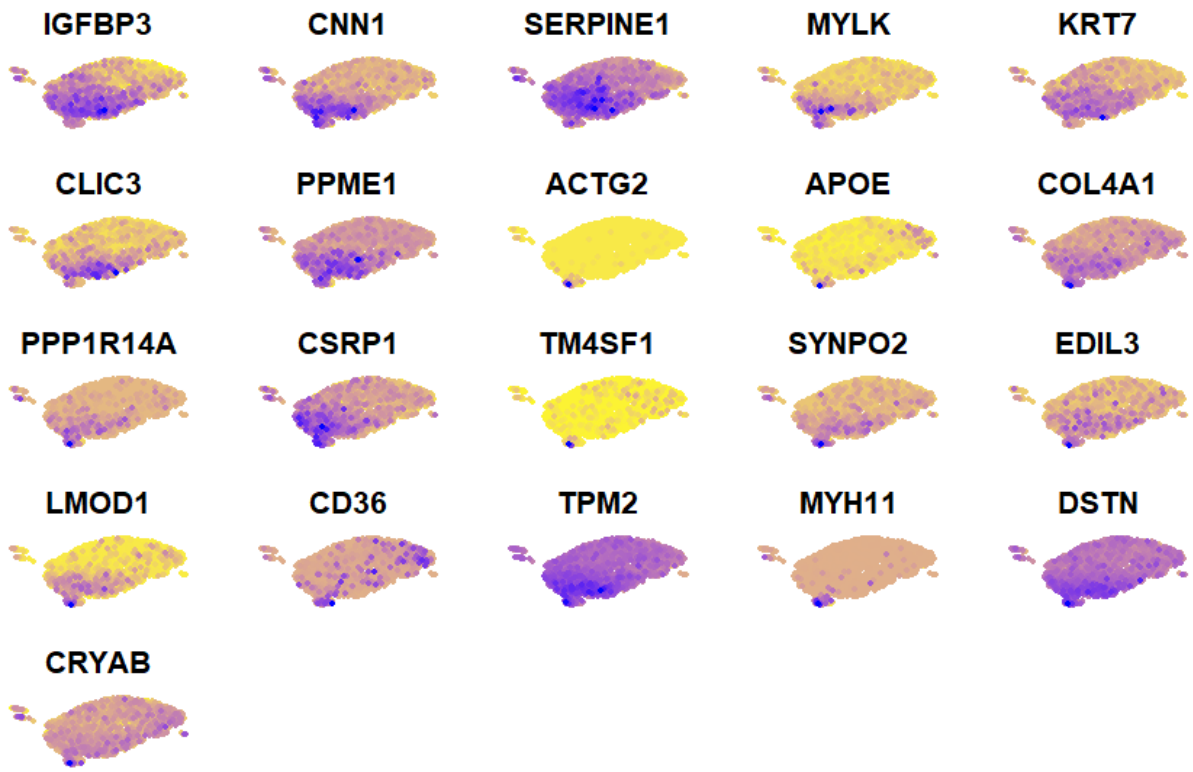


Рисунок 34. Дифференциально экспрессирующиеся гены в α-sma^{+} - субпопуляции

MYLK [242], CNN1 [243], MYL9 [244], PPP1R14A (является ингибитором PPP1CA) участвуют в сокращении гладкомышечных клеток. SYNPO2 [245] участвует в формировании сократительных стрессовых волокон актина, ориентированных параллельно длинной оси клетки. LMOD1 [246] участвует в организации актиновых филаментов (Рис. 34).

Таблица 15. Кластеризация по GO_BP (Gene ontology, biological process) высокоэкспрессированных генов (> 2 раз) в α-sma^{+} - субпопуляции.

GO-term	description	count in network	Padj
GO:0030029	Actin filament-based process	18/824	$1.473*10^{-4}$
GO:0030036	Actin cytoskeleton organization	17/723	$1.258*10^{-4}$
GO:0097435	Supramolecular fiber organization	15/739	$5.248*10^{-3}$
GO:0007015	Actin filament organization	13/449	$5.250*10^{-4}$
GO:0006936	Muscle contraction	10/368	$2.465*10^{-2}$
GO:0110053	Regulation of actin filament organization	9/288	$2.296*10^{-2}$
GO:1902905	Positive regulation of supramolecular fiber organization	8/216	$2.140*10^{-2}$
GO:0051495	Positive regulation of cytoskeleton	8/234	$3.821*10^{-2}$

	organization		
GO:0030239	Myofibril assembly	5/68	$4.142*10^{-2}$

Высокоэкспрессирующиеся гены входят в состав кластеров GO_BP, объединяющих гены, отвечающие за регуляцию организации актиновых волокон, что отражает направление изучаемого нами направления дифференцировки МСК в миофибробласты (табл. 15).

Для того, чтобы в последующем можно было отсортировать нужные клетки, необходимо найти мембранные маркеры. Обычно, для этого используют кластеризацию значимо представленных генов в кластерах.

3.2.7 Мембранные белки α -sma$^-$ - субпопуляции

Из базы данных UniProt были экспортированы таблицы, содержащие информацию о локализации белков в клетке. Сопоставив список наших белков с этой базой, были определены белки, являющиеся мембранными (Табл. 16).

Для сортинга клеток α-sma$^-$ - субпопуляции был выбран один белок, так как он преимущественно распределен именно в данной субпопуляции.

Таблица 16. Мембранные белки клеток субпопуляции α-sma$^-$

Ген	Описание	Кратность изменения	Кластер
CTSK	cathepsin K	2,92	0
MARCKSL1	MARCKS like 1	1,53	0
IL1R1	interleukin 1 receptor type 1	1,47	0
CTSL	cathepsin L	1,46	0
GPNMB	glycoprotein nmb	1,45	0
PDGFRA	platelet derived growth factor receptor alpha	1,36	0
PLPP3	phospholipid phosphatase 3	1,31	0
DPP4	dipeptidyl peptidase 4	1,30	0
ACKR4	atypical chemokine receptor 4	1,29	0
SNX9	sorting nexin 9	1,28	0
EMP2	epithelial membrane protein 2	1,27	0
PDPN	podoplanin	1,25	0
IFITM1	interferon induced transmembrane protein 1	1,18	0

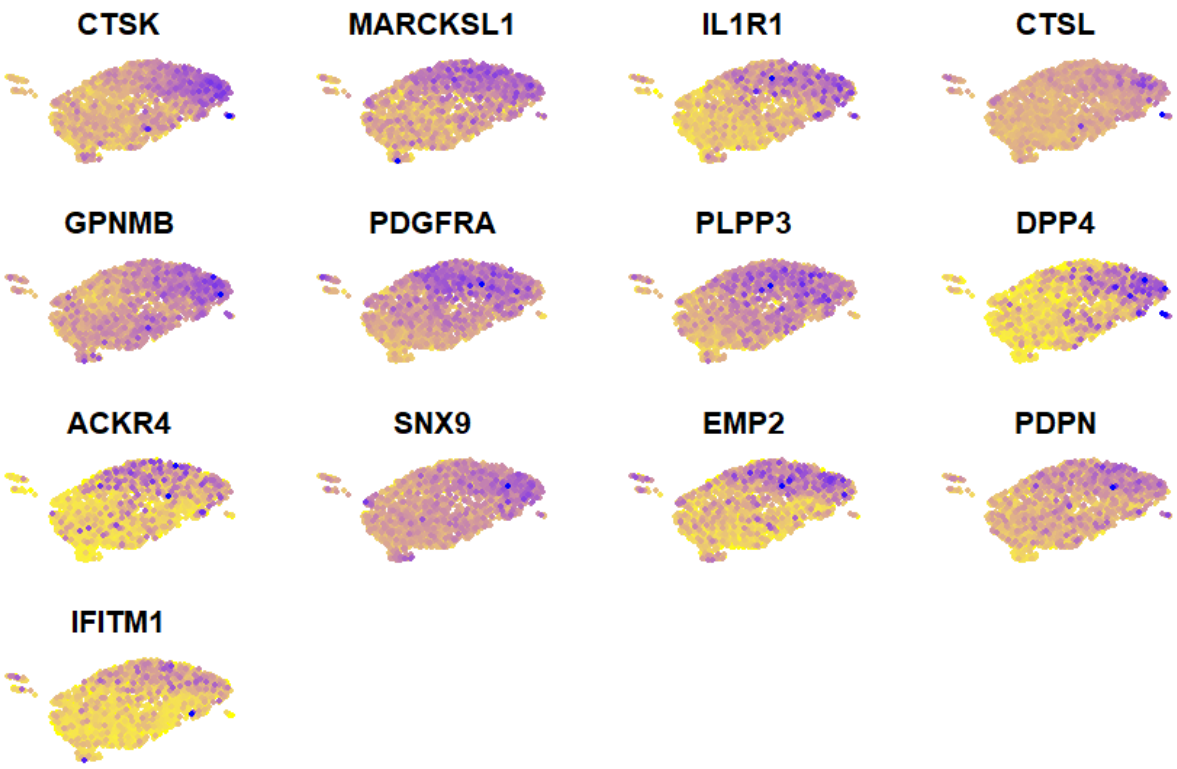


Рисунок 35. Дифференциально экспрессирующиеся гены мембранных белков в α-sma^{-} - субпопуляции

Кроме выбранного мембранного белка были обнаружены и другие, но их распределение не подходит под требования мембранного белка для сортинга (Рис. 35).

3.2.8 Мембранные белки α -sma^{+} - субпопуляции

Для сортинга клеток α-sma^{+} - субпопуляции было отобрано три белка, так как они преимущественно распределены именно в данной субпопуляции (Табл. 17).

Таблица 17. Мембранные белки клеток субпопуляции α-sma^{+}

Ген	Описание	Кратность изменения	Кластер
SLC16A3	solute carrier family 16 member 3	1,19	2
LIMS2	LIM zinc finger domain containing 2	1,33	2
CRIM1	cysteine rich transmembrane BMP regulator 1	1,32	2
CDH2	cadherin 2	1,64	2
CD55	CD55 molecule (Cromer blood group)	1,19	2
CAV2	caveolin 2	1,52	2
CAP1	cyclase associated actin cytoskeleton regulatory protein 1	1,29	2
ACTN1	actinin alpha 1	1,54	2
CAPN2	calpain 2	1,24	2

CAVIN1	caveolae associated protein 1	1,29	2

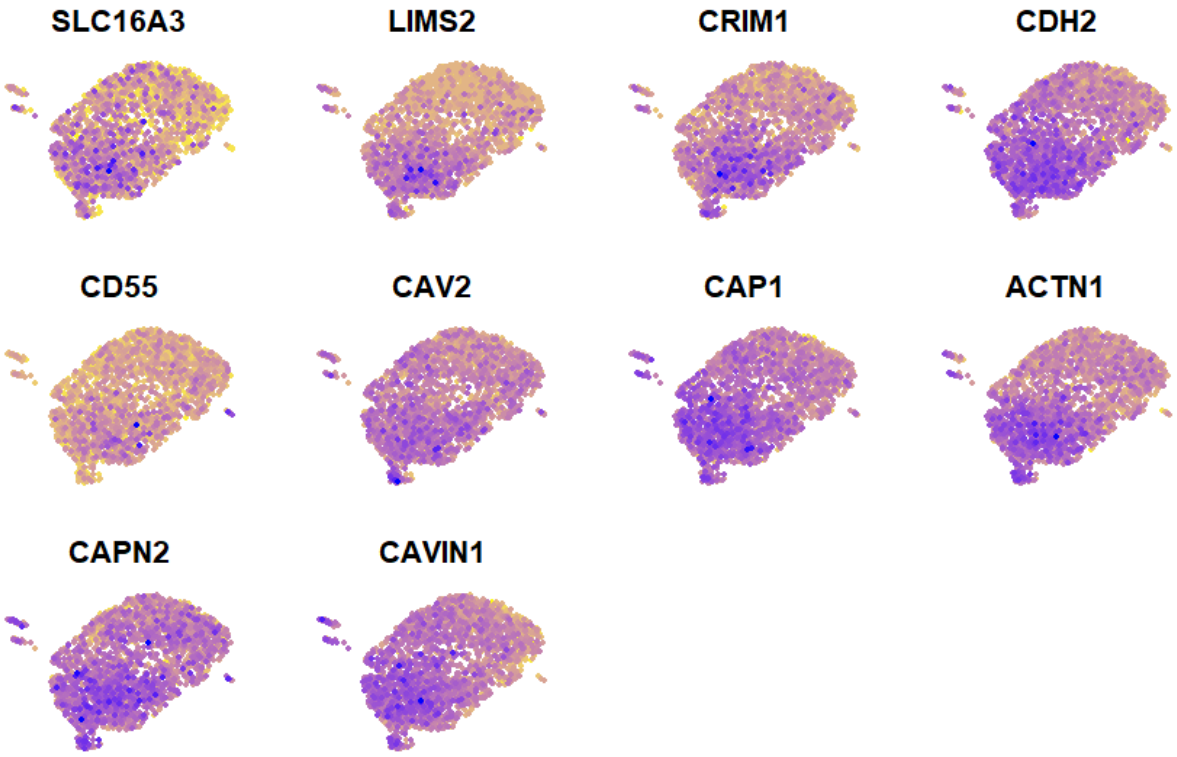


Рисунок 36. Дифференциально экспрессирующиеся гены мембранных белков в α-sma^{+} - субпопуляции

Кроме выбранного мембранного белка были обнаружены и другие, но их распределение не подходит под требования мембранного белка для сортинга (Рис. 36).

3.3 Применение методов типирования клеток образца МСК, культивируемых под влиянием профибротических стимулов

3.3.1 Автоматическое типирование

При использовании референса HumanPrimaryCellAtlasData в контрольном образце определено 7 типов клеток, при использовании BlueprintEncodeData – 6 типов (Рис. 37).

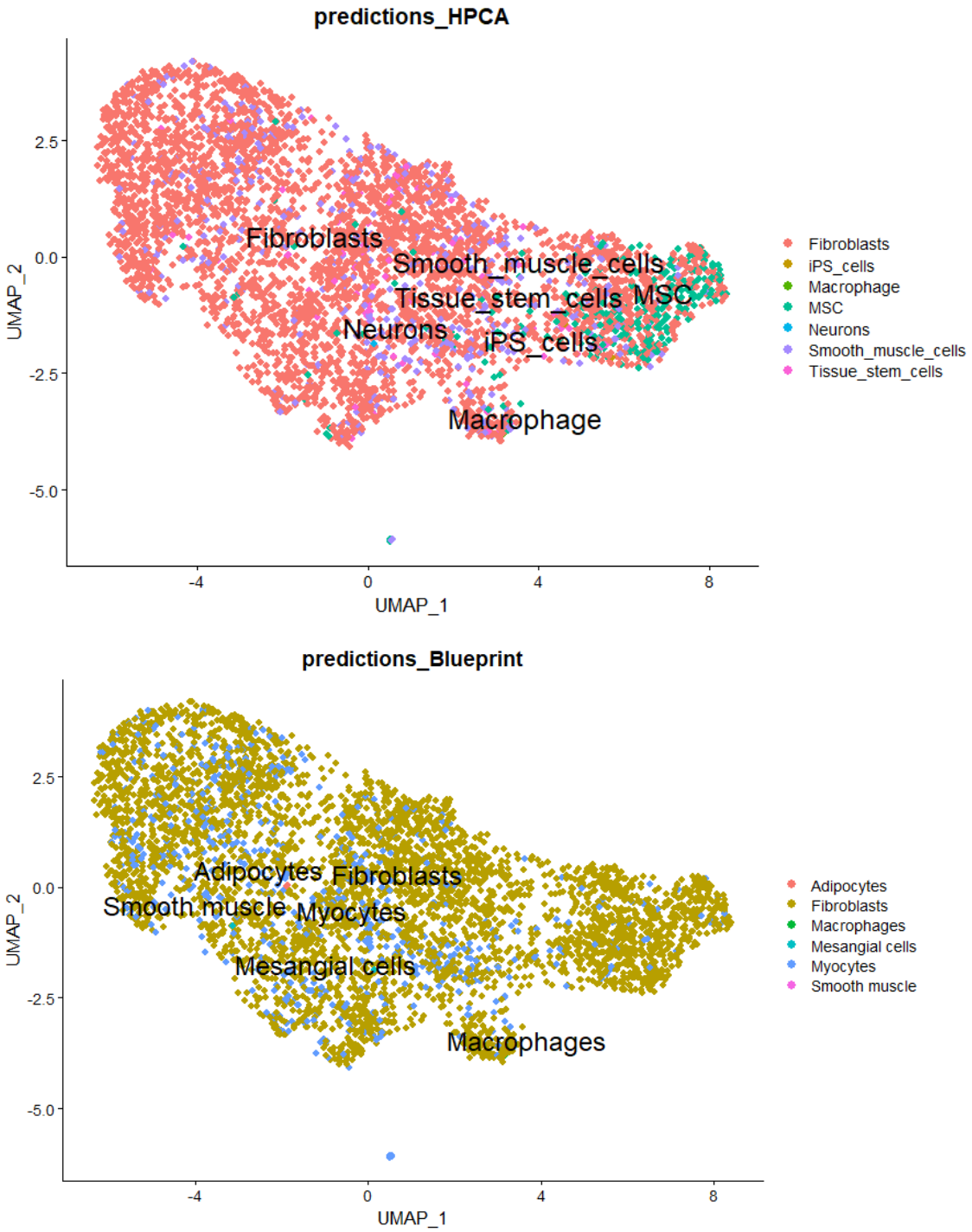


Рисунок 37. Автоматическое типирование клеток контрольного образца с помощью R-пакета SingleR с использованием клеточных референсов HumanPrimaryCellAtlasData и Blueprint

Таблица 18. Количество клеток разных типов в контрольном образце при использовании референсов Blueprint и HPCA

HPCA	Кол-во клеток	Blueprint	Кол-во клеток
Fibroblasts	3668	Adipocytes	2
iPS_cells	3	Fibroblasts	3850
Macrophage	1	Macrophages	1
MSC	272	Mesangial cells	2

Neurons	1	Myocytes	598
Smooth_muscle_cells	416	Smooth muscle	1
Tissue_stem_cells	93		

С помощью автоматического типирования в контрольном образце при использовании клеточного референса HumanPrimaryCellAtlasData было идентифицировано 7 типов клеток: фибробласты, индуцированные стволовые клетки, макрофаги, мезенхимальные стволовые клетки, нейроны, гладкомышечные клетки и тканевые стволовые клетки (Табл. 18). Имея представление о ткани, из которой были выделены клетки нашего образца, можно сказать, что в ткани изначально действительно могли быть все эти клетки, так как в состав подкожной жировой ткани входят все вышеперечисленные клетки. При использовании клеточного референса Blueprint было идентифицировано 6 типов клеток: адипоциты, фибробласты, макрофаги, мезенгиальные клетки, миоциты и гладкомышечные клетки. Эти клетки, также могли быть среди клеток выделенного образца.

При использовании референса HumanPrimaryCellAtlasData в образце МСК, культивируемых в профибротических условиях определено 8 типов клеток, при использовании BlueprintEncodeData – 5 типов (Рис. 38).

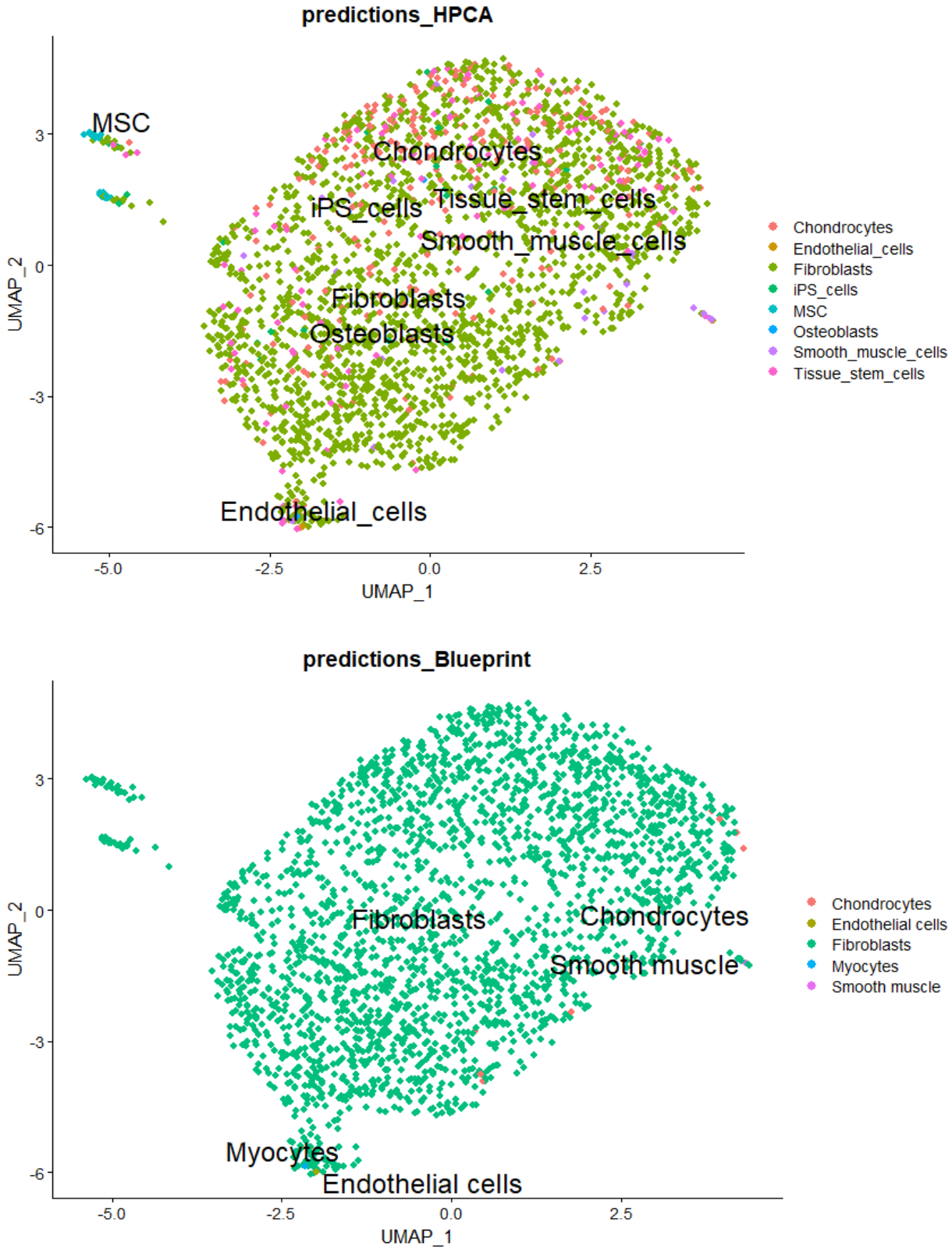


Рисунок 38. Автоматическое типирование МСК, культивируемых в профибротических условиях с помощью R-пакета SingleR с использованием клеточных референсов HumanPrimaryCellAtlasData и Blueprint

Таблица 19. Количество клеток разных типов в образце МКС, культивируемых в профибротических условиях, при использовании референсов Blueprint и HPCA

HPCA	Кол-во клеток	Blueprint	Кол-во клеток
Chondrocytes	269	Chondrocytes	8
Endothelial_cells	1	Endothelial cells	1
Fibroblasts	1919	Fibroblasts	2367
iPS_cells	23	Myocytes	1
MSC	24	Smooth muscle	1
Osteoblasts	2		
Smooth_muscle_cells	29		
Tissue_stem_cells	111		

С помощью автоматического типирования в образце МСК, культивируемых в профибротических условиях при использовании клеточного референса HumanPrimaryCellAtlasData было идентифицировано 8 типов клеток: хондроциты, эндетелиальные клетки, фибробласты, индуцированные стволовые клетки, мезенхимальные стволовые клетки, остеобласты, гладкомышечные клетки и тканевые стволовые клетки (Табл. 19). Если сравнивать результаты автоматического типирования контрольного образца и образца МСК, культивируемых в профибротических условиях, можно отметить, что индуцированные стволовые клетки и макрофаги, определяемые в контрольном образце исчезают в индуцированном образце. В образце МСК, культивируемых в профибротических условиях появляются хондроциты, что, возможно, говорит о том, что под влиянием профибротического окружения и TGFb часть МСК дифференцируются в хондроциты.

При использовании клеточного референса Blueprint было идентифицировано 5 типов клеток: хондроциты, эндотелиальные клетки, фибробласты, миоциты и гладкомышечные клетки. Эти клетки, также могли быть среди клеток выделенного образца.

Автоматическое типирование можно использовать исключительно для ознакомления с массивом данных. Также, автоматическое типирование лучше работает с массивами данных нервных клеток или клеток крови. Для

типирования клеток соединительной ткани лучше ориентироваться на общепринятые специфические маркеры.

3.3.2 Типирование клеток по специфическим маркерам

После кластеризации клеток возникает вопрос о типировании клеток. Обычно, клетки типируют по высокоэкспрессирующимся в кластере специфическим маркерам отдельных типов клеток.

В R-пакете Seurat также есть возможность получения списка дифференциально экспрессирующихся генов с помощью функции **FindAllMarkers**. По умолчанию Seurat выполняет дифференциальное выражение на основе непараметрического критерия суммы рангов Уилкоксона, выводя список наиболее дифференциально экспрессированных генов.

Используя данную функцию, был получен список дифференциально экспрессированных генов для всех кластеров контрольного образца. Учитывая то, что на клетки контрольной популяции еще не оказывалось никакого воздействия, набор клеточных типов в ней может быть обусловлен только составом ткани, из которой выделяли эти клетки. Поскольку клетки выделялись из подкожной жировой ткани, то ожидаемыми для этой ткани могут быть МСК, фибробласты, адипоциты, эндотелиоциты и т.д. Поскольку подготовка библиотеки и секвенирование проводились на 4 день культивирования клеток, можно предположить, что большинство клеток уже запустили какие-то процессы направляющие вектор их развития в сторону запланированного клеточного типа или необходимого состояния. Следовательно, обнаружение в каждом кластере четко различимых клеточных типов будет затруднено. Действительно, из 6 кластеров контрольного образца по специфическим маркерам можно идентифицировать только 2 кластера, как МСК.

Если сравнивать отобранные алгоритмом гены с клеточными референсами Cell Marker и PanglaoDB, они будут идентифицировать

мезенхимальные стволовые клетки. Однако гены, отобранные алгоритмом в других кластерах специфическими не являются и по ним нельзя типировать клетки.

Из результатов видно, что при типировании клеток кластеров образца МСК, культивируемых в профибротических условиях, идентифицируется большее число типов клеток, но это не дает целостного и полного понимания того, как клетки реагируют на стимулы микроокружения, какие процессы протекают в клетках в данный момент и в какова траектория развития этих клеток.

Для решения этого вопроса необходимо рассматривать не только дифференциально экспрессированные гены, специфические для определенных типов клеток, но и выраженные биологические процессы на момент секвенирования.

3.3.3 Типирование промежуточных форм клеток по биологическим процессам

Поскольку большинство клеток в процессе дифференцировки находятся в промежуточных состояниях, их невозможно типировать по специфическим маркерам. Но в процессе дифференцировки в клетке последовательно запускается множество процессов, направляющих ее по пути дифференцировки в определенный тип клеток.

Для типирования клеток по биологическим процессам были использованы базы данных STRING [247] (Табл. 20). В базе данных есть возможность построения интерактома наиболее представленных генов в кластере. В режиме multiple proteins импортируется список высокопредставленных генов. Если применить метод типирования клеток по биологическим процессам, то получится следующий результат.

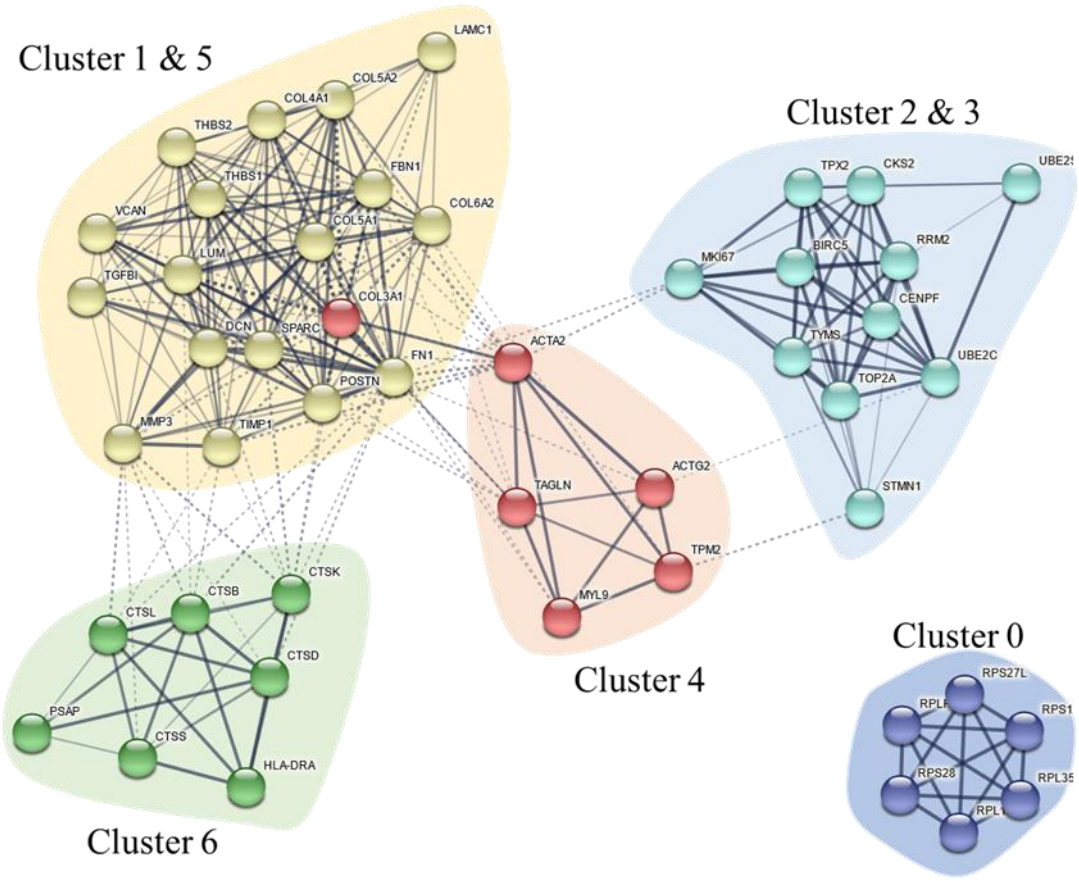


Рисунок 39**.** Интерактом наиболее представленных генов МСК контрольного образца

Таблица 20. Типирование кластеров контрольного образца по биологическим процессам

Cluster	GO:BP	Genes
0	GO:0006412 Translation	RPS12, RPS28, TMA7, THBS1, RPL35A, RPS27L, RPL12, RPLP2
1	Fibroblasts	**MMP3**, SFRP4, PI16
2	Smooth muscle cells	**TIMP1**, **ACTA2**, **TPM2**
3	GO:0000278 Mitotic cell cycle	TOP2A, STMN1, MKI67, CENPF, SMC4, UBE2C, TUBB4B, UBE2S, TYMS, RRM2, CKS2, TPX2, BIRC5, TUBA1B
4	GO:0006936 Muscle contraction	**TPM2**, ACTG2, MYL9, CRYAB, MYL12A, GSTO1
5	GO:0030198 Extracellular matrix organization	TIMP1, POSTN, LOX, SPARC, VCAN, COL3A1, MFAP5, FBN1, LUM, FN1, DCN, CLU
6	GO:0006955 Immune response	CHI3L1, CCL2, CTSD, FUCA1, CTSL, CTSK, HLA-DRA, ANXA1, CXCL8, CLU, PSAP, MDK, CTSB, CTSS

При типировании по специфическим маркерам был типирован только 1 и 2 кластеры. В остальных кластерах типировать клетки по специфическим

маркерам не удалось. Рассматривая биологические процессы в клетках не типированных кластеров видно, что в 0 кластере активны процессы трансляции, в 3 кластере – митоз, в 4 – гены мышечного сокращения, в 5 – организация внеклеточного матрикса и в 6 – иммунный ответ (Рис. 39). По сравнению с типированием клеток по специфическим маркерам, типирование по биологическим процессам дает большее количество информации и позволяет восполнить пробелы первого подхода.

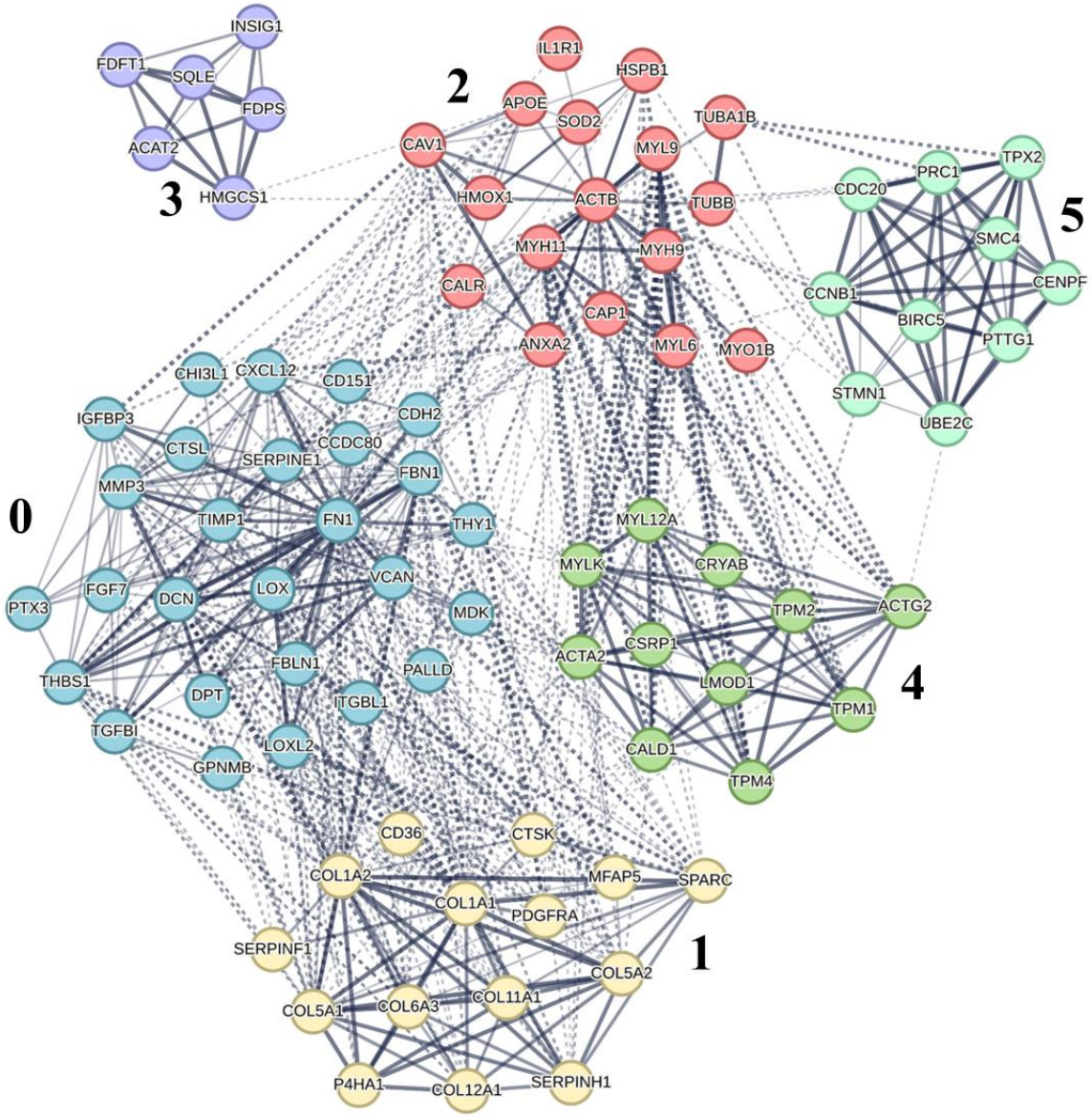


Рисунок 40. Интерактом наиболее представленных генов МСК, культивируемых в профибротических условиях.

Таблица 21. Типирование кластеров по биологическим процессам МСК, культивируемых в профибротических условиях.

Cluster	GO:BP	Genes
0	GO:0030334 Regulation of cell migration	DCN, HMOX1, SERPINF1, PDGFRA, FGF7, FBLN1, GPNMB, CXCL12, MDK, IL1R1, SOD2
	GO:0030198 Extracellular matrix organization	DCN, PDGFRA, CTSK, COL6A3, MMP3, FBLN1, CTSL, DPT
	GO:1901342 Regulation of vasculature development	DCN, HMOX1, SERPINF1, CHI3L1, GPNMB, MDK, SOD2
1	GO:0030198 Extracellular matrix organization	CCDC80, TIMP1, SERPINE1, COL1A1, LOX, SPARC, THBS1, P4HA1, VCAN, PTX3, COL1A2, COL12A1, FBN1, ANXA2, FN1, MFAP5, COL11A1, COL5A1, COL5A2, LOXL2, TGFBI, SERPINH1
	GO:0001568 Blood vessel development	MYH9, SERPINE1, COL1A1, LOX, THBS1, CDH2, THY1, COL1A2, ANXA2, FN1, COL5A1, LOXL2, TGFBI
	GO:0048514 Blood vessel morphogenesis	MYH9, SERPINE1, LOX, THBS1, CDH2, THY1, ANXA2, FN1, LOXL2, TGFBI
	GO:0072359 Circulatory system development	MYH9, SERPINE1, COL1A1, LOX, SPARC, THBS1, CDH2, THY1, COL1A2, CALR, FBN1, ANXA2, FN1, TPM1, COL11A1, COL5A1, LOXL2, TGFBI
	GO:0001503 Ossification	COL1A1, LOX, SPARC, VCAN, COL1A2, TPM4, COL11A1, COL5A2, IGFBP3
	GO:0001501 Skeletal system development	MYH9, COL1A1, LOX, THBS1, VCAN, CDH2, COL1A2, TPM4, FN1, TPM1, COL5A1, CAP1, ITGBL1, LOXL2, MYO1B, PALLD
2	GO:0006936 Muscle contraction	MYL12A, MYL9, TPM1, MYLK, CALD1, TPM2, ACTA2, MYL6
	GO:0010757 Negative regulation of plasminogen activation	SERPINE1, THBS1
3	GO:0006695 Cholesterol biosynthetic process	SQLE, HMGCS1, INSIG1, FDPS, ACAT2, FDFT1
4	GO:0006936 Muscle contraction	MYL9, TPM1, MYLK, CALD1, LMOD1, TPM2, MYH11, ACTG2, ACTA2, CRYAB, MYL6
	GO:0042060 Wound healing	MYH9, HSPB1, MYL9, ACTB, TPM1, CSRP1, CD151, CD36
	GO:0061041 Regulation of wound healing	MYH9, APOE, CAV1, MYLK, CD36
5	GO:0051301 Cell division	CCNB1, TPX2, BIRC5, TUBA1B, TUBB, UBE2C, SMC4, CENPF, CDC20, PTTG1, PRC1, STMN1

Рассматривая биологические процессы в клетках не типированных кластеров видно, что в 0 кластере выражены процессы, связанные регуляцией миграции, организацией внеклеточного матрикса и регуляции развития сосудов, в 1 кластере – организация внеклеточного матрикса, развитие кровеносной системы, развитие опорно-двигательной системы, во 2 – мышечное сокращение и отрицательная регуляция активации плазминогена, в 3 – биосинтез холестерола, в 4 – мышечное сокращение и заживление ран и в 5 - клеточное деление (Рис. 40). По сравнению с типированием клеток по специфическим маркерам, типирование по биологическим процессам дает большее количество информации и позволяет восполнить пробелы первого подхода (Табл. 21).

3.3.4 Типирование клеток, основанное на знании о дифферонах и положении не типированного кластера на траектории развития

Метод типирования клеток основан на знаниях о дифферонах клеток и литературных данных, описывающих морфологические свойства и транскриптом начальных, промежуточных и конечных форм клеток, связанных преемственной линией дифференцировки. Рассматривая влияние модели, воспроизводящей профибротические условия можно сказать, что мезенхимальные стволовые клетки могут дифференцироваться в нескольких направлениях и одним из них является направление МСК → мифибробласт → фибробласт. В таком случае, если при анализе удается с высокой долей вероятности типировать мезенхимальные стволовые клетки и фибробласты, а в кластере находящемся на траектории между ними специфических маркеров не определяется, то можно предположить, что в промежуточном кластере, на основании знаний о дифферонах, находятся миофибробласты.

При анализе RNA-velocity векторов МСК, культивируемых в стандартных условиях, не было обнаружено нетипированных кластеров, находящихся в промежуточном положении между кластерами с известными типами клеток (Рис. 41).

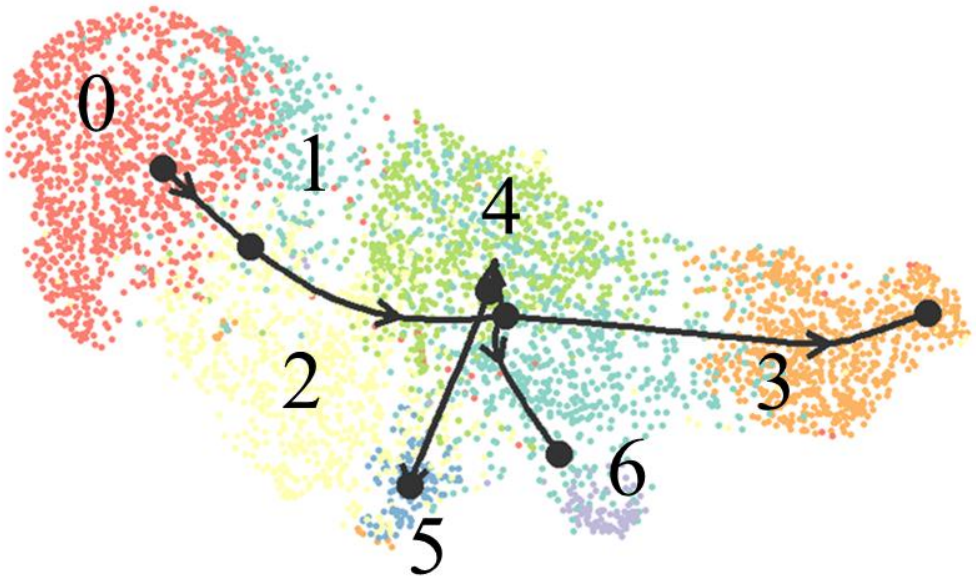


Рисунок 41. Траектория развития клеток контрольного образца. 0 - GO:0006412 Translation; 1 - Fibroblasts (MMP3); 2 - Smooth muscle cells (TIMP1, ACTA2, TPM2); 3 - GO:0000278 Mitotic cell cycle; 4 - Fibroblasts (TPM2); 5 - GO:0030198 Extracellular matrix organization; 6 - GO:0006955 Immune response

При анализе RNA-velocity векторов МСК, культивируемых в профибротических условиях, выявлен 2 кластер, находящийся на траекториях, начинающихся в 0, 1 и 3 кластерах и заканчивающихся в 4 кластере. В 0 кластере располагаются гладкомышечные клетки, в 1 остеокласты, в 3 у клеток нет специфического маркера, но известно, что в клетках данного кластера активно выражены антиоксидативные процессы, фибринолиз и адипогенная дифференцировка. В 4 кластере типированы МСК. Поскольку 2 кластер располагается на траектории, начинающейся из 1 кластера с остеокластами к 4 кластеру с МСК и 0 кластеру с гладкомышечными клетками, можно предположить, что в нем располагаются фибробласты, направляющиеся в дедифференцировку в МСК с последующей дифференцировкой в гладкомышечные клетки (Рис. 42).

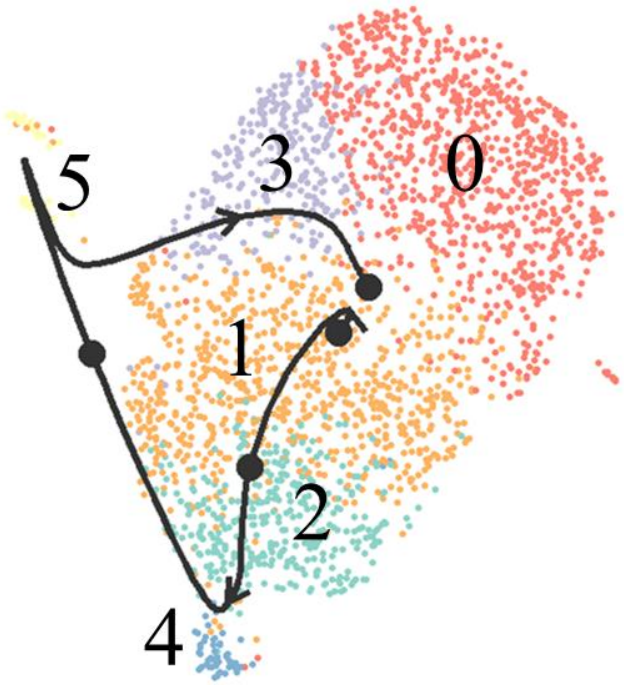


Рисунок 42. Траектория развития МСК, культивируемых в профибротических условиях. 0 - Smooth muscle cells (ACTA2) GO:0030198 Extracellular matrix organization & GO:1901342 Regulation of vasculature development; 1 - GO:0030198 Extracellular matrix organization & GO:0048514 Blood vessel morphogenesis; 2 - GO:0006936 Muscle contraction & GO:0010757 Negative regulation of plasminogen activation; 3 - GO:0006695 Cholesterol biosynthetic process; 4 - GO:0006936 Muscle contraction & GO:0042060 Wound healing; 5 - GO:0051301 Cell division

3.4 Направления развития МСК, культивируемых в профибротических условиях

Интеграция проводится для идентификации субпопуляций клеток, присутствующих в обоих образцах, получения списка клеточных маркеров присутствующих в обоих образцах и сравнения образцов с целью получения информации об ответе клеток на предъявляемые им стимулы.

Для анализа направлений развития и дифференцировки клеток недостаточно анализа отдельных образцов, так как необходимо провести сравнение клеток обоих образцов друг с другом для изучения кластеров клеток, не пересекающихся между образцами.

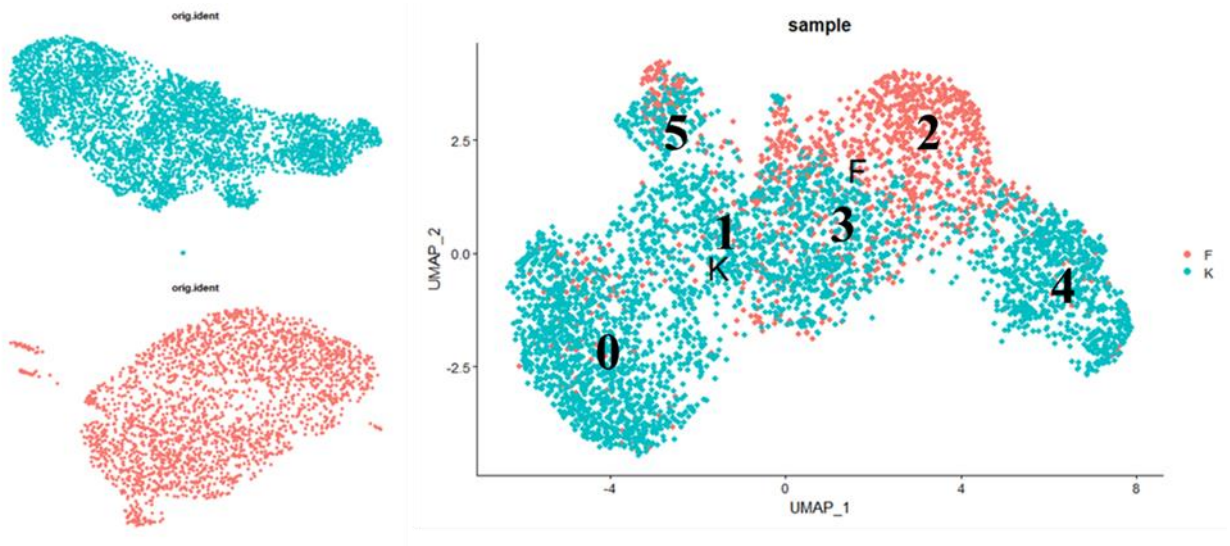


Рисунок 43. Создание интегрированного объекта

При визуализации интегрированного объекта по образцам видно, что клетки контрольного образца и образца МСК, культивируемых в профибротических условиях, перекрываются только в 1 и 3 кластерах. 2 и 5 кластеры представляют особый интерес, так как это кластеры клеток, которых не было в объектах отдельных образцов (Рис. 43).

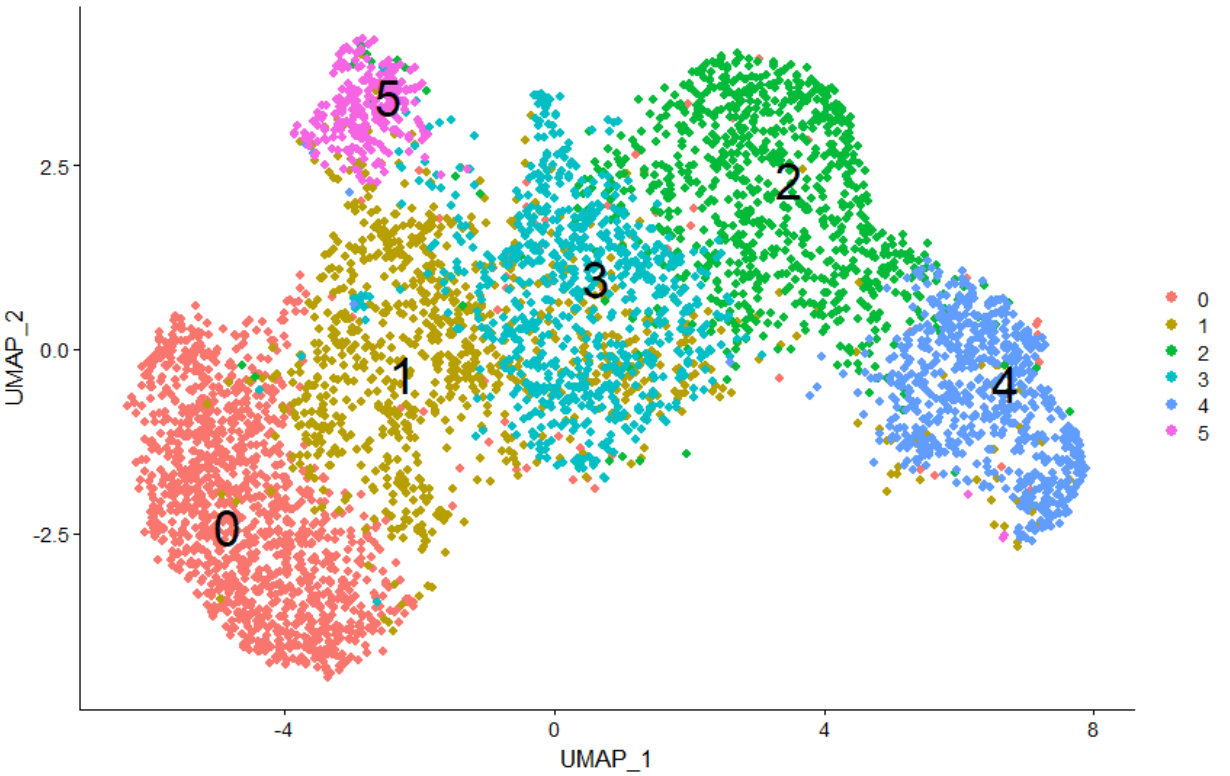


Рисунок 44. UMAP-график кластеризации интегрированного объекта

Таблица 22. Идентифицированные клеточные типы в кластерах интегрированного объекта

Cl	Type	Genes
0	Unknow	LRRC75A, C1orf56, SNHG29, FTH1, OST4, FTL, ATP5F1E, SERF2, B4GALT1
1	Fibroblasts, myofibroblasts	ACTA2, FUCA1, TPM2, TAGLN, CALD1, CTSD, CD36, CCL2, SOD2
2	Fibroblasts	MGP, CLU, DPT, RARRES2, SERPINF1, CTSK, NR4A1, FBLN1, DCN, CYP1B1
3	Basal cells, fibroblasts	ACTA2, TPM2, TAGLN, KRT7, CDKN1A, NUPR1, MYL9, CLIC3, PPME1
4	Germ cells	HIST1H4C, TOP2A, H2AFZ, DEK, MKI67, CENPF, HMGB2, RRM2, UBE2C
5	Fibroblasts	NEAT1, TIMP1, POSTN, LOX, HSPA5, TIMP3, DKK1, VCAN, COL3A1

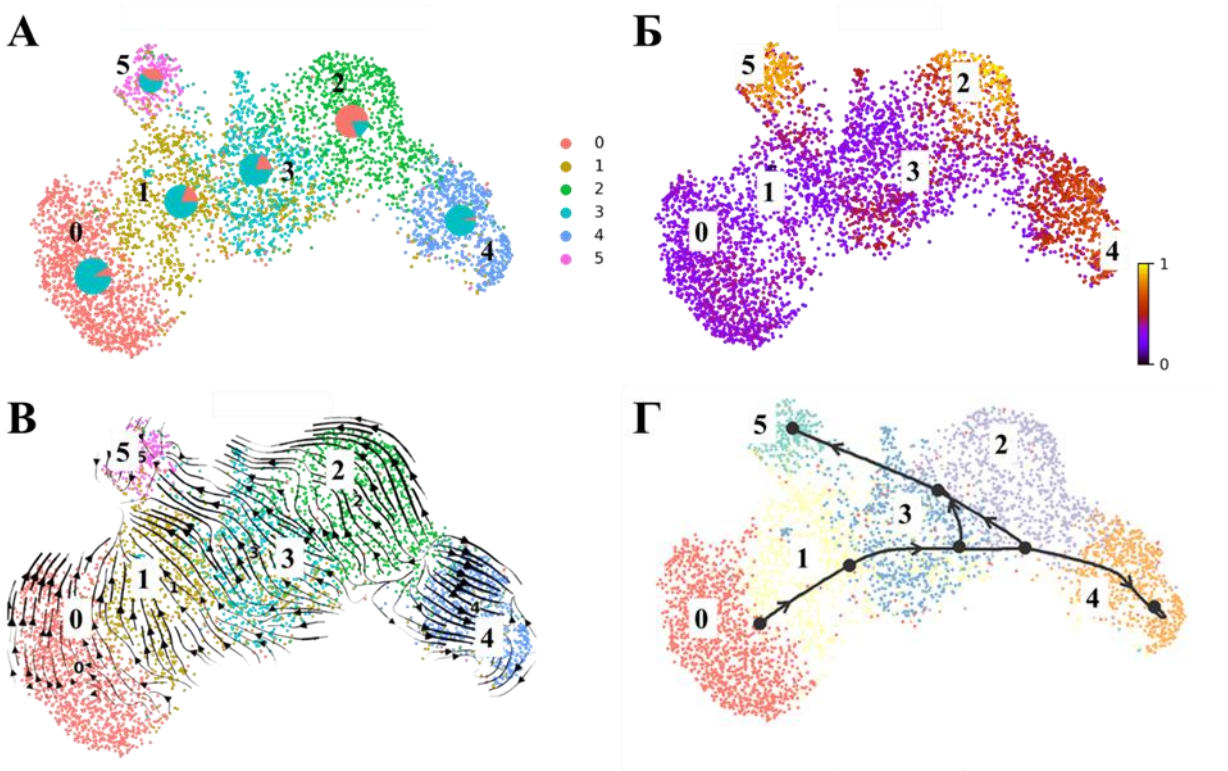


Рисунок 45. Downstream анализ клеток интегрированного объекта. А. Результат кластеризации клеток интегрированного объекта с PiePlot, обозначающими соотношение клеток контрольного и профибротического образцов в каждом кластере (контроль – коралловый, фиброз - голубой); Б. График латентного времени клеток интегрированного объекта (положение клеток на временной линии дифференцировки или развития); В. Векторы RNA-velocity клеток интегрированного объекта, показывающие направления дифференцировки клеток внутри кластеров; Г. Paga plot с траекториями

развития клеток интегрированного образца, показывающими направления дифференцировки клеток между макросостояниями (кластерами).

Изучение траекторий развития клеток дает значительные преимущества, так как можно анализировать не только дифференциальную экспрессию, но также появляется возможность проследить направление развития или дифференцировки клеток в ответ на стимулы или окружение.

Так, одна из траекторий в интегрированном образце (Рис. 44,45, Г; Табл. 22) начинается с клеток, представленных в 0 кластере, где определение клеточного типа отсутствует, что говорит о наличии стволовых или недифференцированных клеток, готовых к дальнейшему развитию. Затем клетки направляются к кластеру 1, где присутствуют миофибробласты, отвечающие за сокращение гладкой мускулатуры (HSA-445355). Эти клетки обладают высокой пластичностью и могут трансформироваться в другие типы клеток в ответ на сигналы окружающей среды, что связано с процессами регенерации и заживления. После этого клетки проходят через кластер 3 с фибробластами и базальными клетками, которые активно участвуют в регуляции клеточной пролиферации и формировании внеклеточного матрикса (GO:0042127, GO:0005201). Наконец, траектория завершается в 4 кластере, состоящем из активно делящихся клеток, задействованных в клеточном делении (GO:0051301) и митотическом цикле (GO:0000278). Это конечное состояние указывает на активный рост и восстановление ткани, обеспечивая жизнеспособность клеточной популяции.

Вторая траектория, начинающаяся во 2 кластере с фибробластами, также проходит через базальные клетки в 3 кластере и завершает путь в 5 кластере с фибробластами и базальными клетками. Эта траектория акцентирует внимание на процессах организации коллагена (GO:0030199), развития соединительной ткани (GO:0061448), заживления ран (GO:0042060) и негативной регуляции активности металлопротеиназ (GO:1905049). Тканевая регенерация и восстановление зависят от организованной работы

внеклеточной матрицы, что предоставляет структурную поддержку и способствует восстановлению поврежденных участков.

График RNA-velocity (Рис. 45, В) повторяет основные направления траекторий с уточнениями направлений внутри кластеров. Дополнительные графики с указанием соотношения количества клеток разных образцов в каждом кластере и график латентного времени показывают, что наибольшее изменение пропорций клеток происходит в кластерах 2 и 5 (Рис. 45, А), которые на графике латентного времени имеют наиболее желтый оттенок (Рис. 45, Б), что говорит о том, что в данных кластерах находятся наиболее дифференцированные формы клеток.

При построении траекторий развития для образца МСК, культивируемых в профибротических условиях во втором кластере помимо специфического маркера миофибробластов активны гены, отвечающие за рост сосудов. Было предположено, что это вполне естественная реакция на повреждение. Процесс заживления ран предполагает стимуляцию роста сосудов, по которым в место повреждения доставляются клетки и строительный материал.

Получив данные кластеризации результатов scRNA-seq был сделан вывод, что такие результаты можно использовать в клинике для изучения склонности пациента к фиброзированию тканей.

3.5 Вклад нкРНК в формирование устойчивости к развитию фиброза

В результате анализа генов в соответствии с принадлежностью к кластерам терминов и путей GO, DO, KEGG и Reactome, связанных с фибротическими состояниями, было отобрано 16 генов со сниженной и 9 генов с повышенной экспрессией. Были найдены все микроРНК, для которых данные гены являются мишенями. Полученные микроРНК были поделены на 3 группы: downregulated - 254 микроРНК против группы генов со сниженной экспрессией, upregulated - 127 микроРНК против группы генов с повышенной экспрессией, overlapped - 51 микроРНК, имеющие мишенями как гены из

группы сниженной экспрессией, так и гены из группы повышенной экспрессии. Были подсчитаны общее количество мишеней каждой микроРНК, а также их специфичность по отношению к отобранным генам (Рис. 46, 47). Медианные значения специфичности составили 0,8% для групп upregulated и downregulated, и 0,4% как для мишеней с повышенной, так и пониженной экспрессией в группе overlapped. Медианные значения по числу мишеней составляют 150 мишеней для групп downregulated и upregulated, и примерно 300 мишеней для группы overlapped.

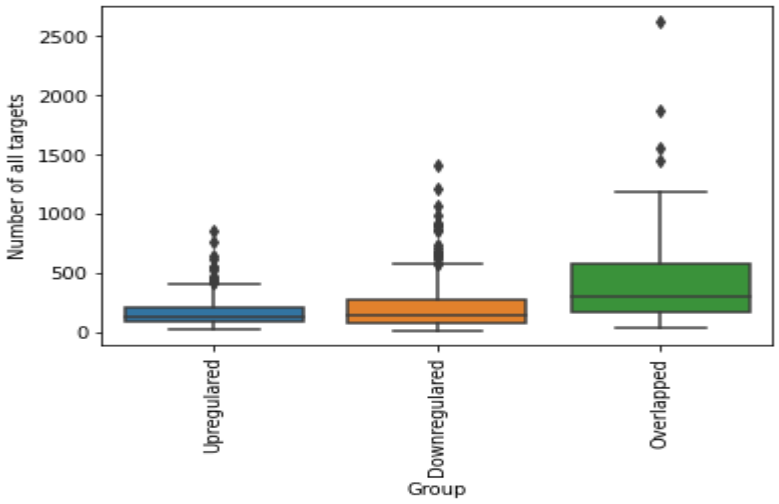


Рисунок 46. Бокс-плот, отображающий количество мишеней микроРНК

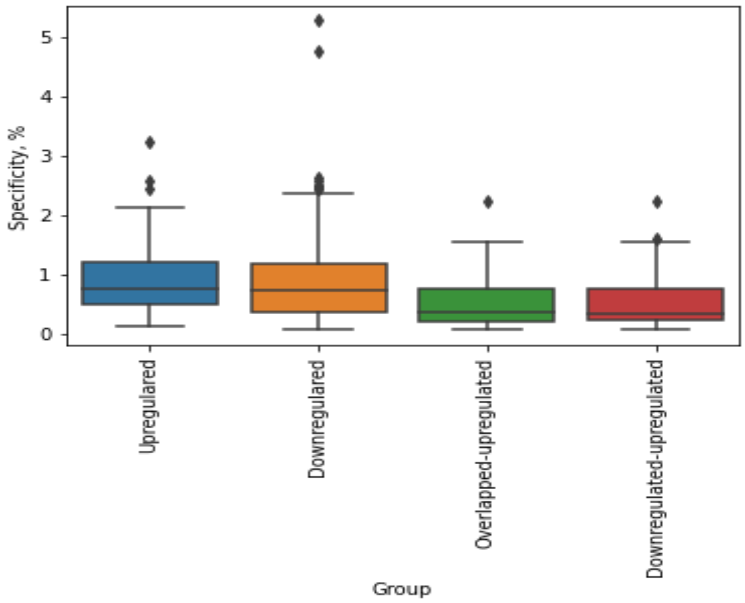


Рисунок 47. Бокс-плот, отображающий процент специфичности микроРНК к отобранным генам-мишеням

Для сужения круга миРНК был проведен семантический анализа и обогащения их генов-мишеней с фильтрацией по терминам, связанным с процессами регуляции процесса фиброза и МСК ('stem', 'fibrosis', 'wound', 'differentitation', 'actin', 'matrix', 'fibroblast', 'myofibroblast', 'aging', 'senescence', 'apoptosis', 'wounding', 'muscle', 'programmed', 'extracellular', 'encapsulating', 'proliferation', 'mesenchymal'). Дополнительно был проведен анализа литературы для формирования окончательного списка в кандидатных микроРНК для изучения их влияния на субпопуляции МСК: hsa-mir-10a-5p, hsa-mir-34a-5p, hsa-mir-27a-3p, hsa-mir-194-5p, hsa-mir-18a-5p, hsa-mir-20a-5p, hsa-mir-451a, hsa-mir-129-5p, hsa-mir-29a-3p, hsa-mir-29b-3p, hsa-mir-29c-3p.

Благодаря возможности кастомизации референсного генома, после добавления генов нкРНК было найдено несколько днРНК, из которых LINC01705 дифференциально экспрессировалась в 0 кластере, который относится к субпопуляции клеток, где происходит сдерживание фиброза.

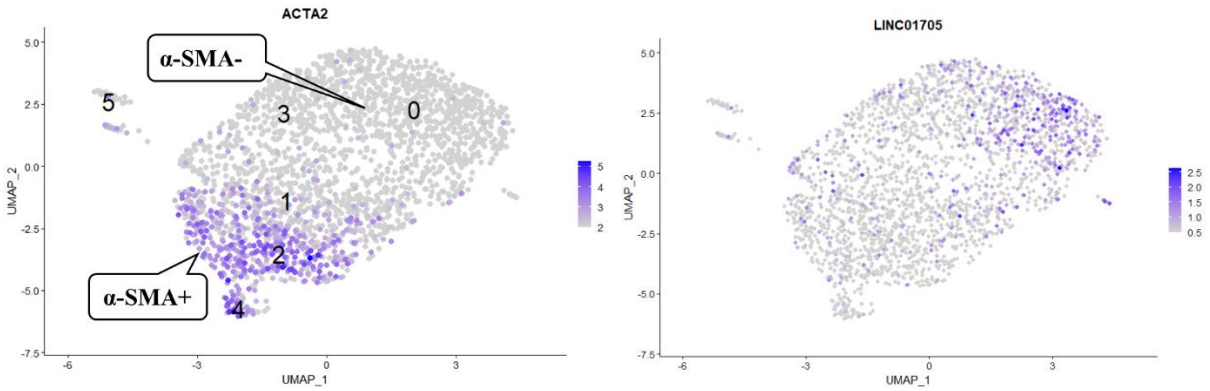


Рисунок 48. Распределение днРНК LINC01705 в клетках образца МСК, культивируемых под воздействием профибротических условий.

Среди некодирующих РНК в антифибротической субпопуляции была выявлена LINC01705 (Рис. 48). По данным ENSEMBLE [248] данная последовательность относится к длинным некодирующим РНК (днРНК), имеет 3 транскрипта: LINC01705-201 (562 пн.), LINC01705-202 (1070 пн.), LINC01705-203 (863 пн.). По классификации относительно белок кодирующих генов относится к межгенным днРНК. Интересно, что по

данным базы данных LncExpDB [249] эта РНК чаще всего встречается в мезенхимных стромальных клетках легких и яичек.

Анализ экспрессии, с помощью базы данных Expression Atlas, показал, что LINC01705, имеет низкую экспрессию (не выше 3 TPM) в 14 здоровых тканях, проявляет наибольшую экспрессию в choroid plexus (сосудистое сплетение желудочков мозга). Однако более высокую экспрессию (более 12 TPM) LINC01705 проявляет в экспериментах, связанных с раковыми заболеваниями поджелудочной железы.

Анализ экспрессии для каждого транскрипта с помощью программы LocExpress [250] показал, все транскрипты экспрессируются в идентичных тканях: селезенке, предстательной железе, легких, печени, почках. Однако LINC01705-202 также экспрессируется в крови и скелетных мышцах.

Анализ экспрессии транскриптов в экспериментах, связанных с раковыми заболеваниями, показал более высокий уровень экспрессии. Таким образом транскрипты экспрессируются в заболеваниях кожи, яичника, предстательной железы и лейкемии плазмы клеток (skin cutaneous melanoma, prostate adenocarcinoma, ovarian serous cystadenocarcinoma, multiple myeloma plasma cell leukemia).

Ab initio предсказание взаимодействия с белками для каждого транскрипта показало, что LINC01705-201 взаимодействует с 73 белками. Наилучшее показатели p-value < 0.001 продемонстрировали белки FLG, SRRM2 и ZNF469. LINC01705-202 - 96 белками. Наилучшее показатели p-value < 0.001 продемонстрировали белки ZNF469, PCF11 и NCL. LINC01705-203 - 716 белками. Наилучшее показатели p-value < 0.001 продемонстрировали белки ZNF469, FLG и SPEN.

Таким образом белок ZNF469 (zinc finger protein 469) проявляет лучшее взаимодействие со всеми транскриптами. По данным NCBI данный белок отвечает за синтез волокон коллагена. Мутации данного ZNF469 могут вызывать brittle cornea syndrome.

Также белок FLG (filaggrin) проявляет взаимодействие с двумя транскриптами. По данным UniProt данный белок агрегирует кератиновые промежуточные филаменты и способствует образованию дисульфидных связей между промежуточными филаментами во время терминальной дифференцировки эпидермиса млекопитающих.

Для оценки возможности регуляции процессов, связанных с фиброзом, был получен список микроРНК, с которыми данная днРНК может взаимодействовать. После этого среди генов-мишеней каждой из отобранных 13 микроРНК было получено количество генов, относящихся к одному из терминов GO: extracellular structure organization, response to wounding, tissue morphogenesis, tissue remodeling, wound healing. У hsa-mir-335-5p больше всего мишеней в 3 из 5 перечисленных процессов, а процессы tissue remodeling, wound healing управляются наибольшим количеством микроРНК. Среди генов, участвующих в регуляции фиброза и таргетируемых наибольшим количеством микроРНК - VEGFA, RTN2, POLM, GATAD1, SPRED1. Некоторые из 13 микроРНК связаны с такими заболеваниями, как цирроз печени и фиброз желчного пузыря.

Для определения ассоциации исследуемой днРНК с заболеваниями применен метод полногеномного поиска ассоциаций [251] (GWAS). По данным базы данных GWAS catalog, LINC01705 ассоциируется с 23 вариантами, которым соответствуют 9 заболеваниям. Исследуемая днРНК наиболее ассоциируется с развитием келоидных рубцов в 3 исследованиях, p-value<10-20. Найденные исследователями варианты картируются на локус гена LINC01705.

Кроме этой информации есть еще результаты исследований в других работах. Недавно была выявлена функция LINC01705 как эндогенной конкурирующей miR-186-5p, регулирующей экспрессию белка TRP при раке груди [252]. В одной из работ приведен список возможных основных регуляторов эпителиально-мезенхимального перехода и TGFβ сигнального пути при протоковой аденокарциноме поджелудочной железы: TGFB2-AS1,

AL138930.1, LINC01705, AC245041.1, UCA1, и NKILA [253]. Предполагают участие lncRNA-LINC01705-201 (ENSG00000232679.2, ENST00000438158.1) в эпителиально-мезенхимальном переходе пигментного эпителия сетчатки [254]. Nanus и соавторы сообщили о 19 дифференциально экспрессирующихся днРНК в фибробластах при сравнении пациентов с остеоартрозом с нормальным весом и пациентами без остеоартроза: MALAT1, MIR155HG, SMILR, LINC01426, RP11-863P13.3, CARMN, RP11-79H23.3, RP11-362F19.1, RP11-290 M5.4, VLDLR-AS1, RP11-536 K7.3, HAGLR, LINC01915, RP11-367F23.2, RP11-392O17.1, LINC01705, LINC01021, DNAJC27-AS1, и AF131217.1 [255].

Для трех транскриптов рассчитана коэкспрессия для раковых заболеваний. Функциональная аннотация дает возможность сделать допущение что LINC01705-202 вовлечен в биологический процесс протеолиз, LINC01705-203 - ферментацию. Для LINC01705-201 не удалось выделить биологический процесс.

3.6 Вклад транскрипционных факторов в подавление фиброза

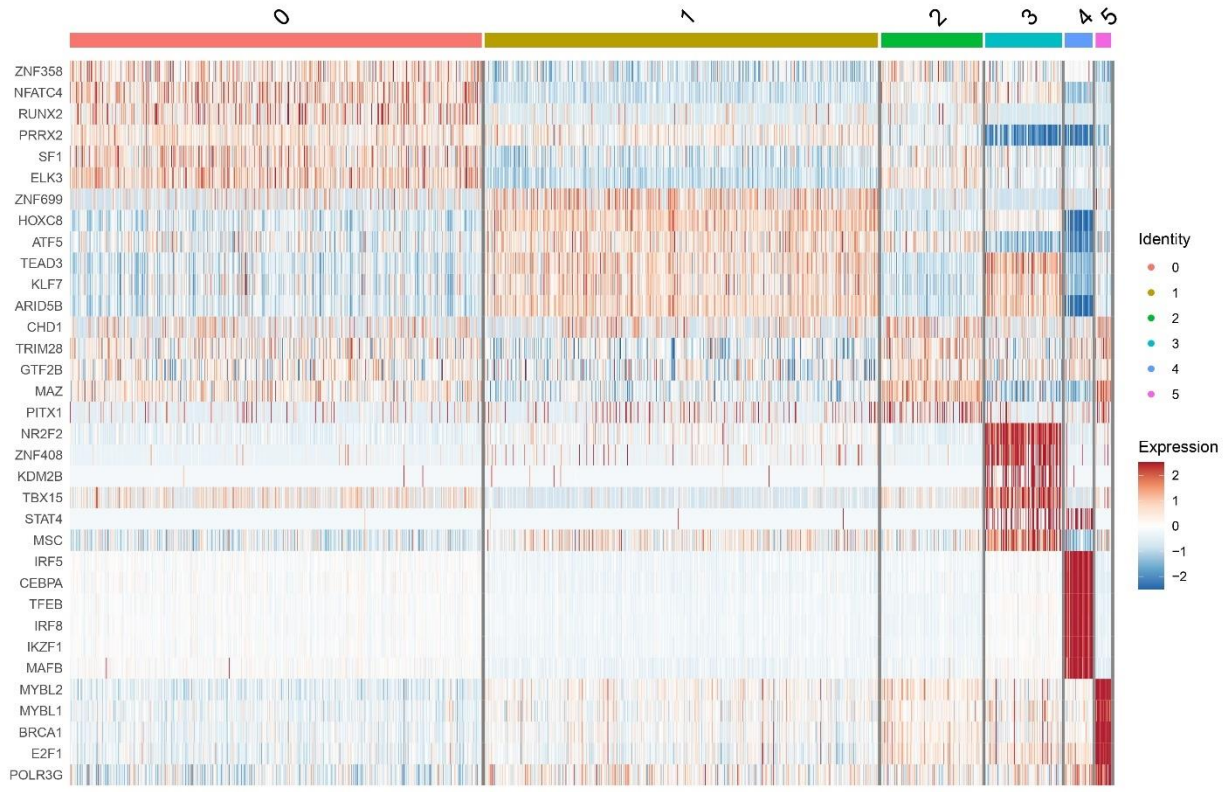


Рисунок 49. Регулоны в кластерах образца МСК, культивируемых в профибротических условиях

Регулонами называется группа генов, регулируемых и коэкспрессирующихся с определенным траснкрипционным фактором (далее – ТФ).

Из транскрипционных факторов регулонов 0 кластера ELK3 является негативным регулятором транскрипции, PRRX2 играет роль в безрубцовом заживлении ран, RUNX2 участвует в остеобластной дифференцировке и формировании скелета, NFATC4 принимает участие в адипогенной дифференцировке и формировании мышечных волокон. Из регулонов 1 кластера ARID5B участвует в адипогенезе, ATF5 в выживании клеток, их пролиферации и дифференцировке. Из регулонов 2 кластера среди генов регулона 2 кластера отсутствуют гены, участвующие в регуляции процесса фиброза. Из регулонов 3 кластера TBX5 контролирует количество предшественников мезенхимальных клеток и хондроцитов. Из регулонов 4

кластера CEBPA участвует в дифференцировке адипоцитов. Из регулонов 5 кластера E2F1 останавливает дифференцировку адипоцитов (Рис. 49).

Для того, чтобы понять, в каких процессах участвуют гены, регулируемые различными ТФ, были получены списки этих генов для каждого ТФ. После этого для каждой группы генов была выполнена кластеризация по GO. Поскольку изначальной задачей было описать антифибротические свойства кластеров, где расположились клетки с низкой экспрессией гладкомышечного актина, для дальнейшего анализа были выбраны кластеры 1 и 3.

В 1 кластере карты белок-белковых взаимодействий были построены только для ТФ TEAD3 и ARID5B, так как количество регулируемых ими генов достаточно для построения графа.

Таблица 23. Биологические процессы для генов ТФ TEAD3

GO-term	description	Count in network	strength	False discovery rate
GO:0030029	Actin filament-based process	10/592	0,78	0,0093
GO:0030036	Actin cytoskeleton organization	9/516	0,79	0,0128
GO:0000904	Cell morphogenesis involved in differentation	9/566	0,75	0,0193
GO:0060429	Epithelium development	12/1109	0,59	0,0294
GO:0098609	Cell-cell adhesion	8/505	0,75	0,0422

Так, для ТФ TEAD3 были выбраны процессы, которые потенциально могут влиять на течение процесса фиброза. Сюда относятся сборка актиновых нитей и организация цитоскелета, дифференцировка клеток, развитие эпителия и межклеточные взаимодействия (Табл. 23).

Таблица 24. Биологические процессы для генов ТФ ARID5B

GO-term	description	Count in network	strength	False discovery rate
GO:0048585	Negative regulation of response to stimulus	37/1636	0.27	0.0143
GO:0045596	Negative regulation of cell differentiation	21/728	0.38	0.0184
GO:0031327	Negative regulation of cellular biosynthetic process	35/1592	0.26	0.0312

Для ТФ ARID5B было отобрано три процесса. Это отрицательная регуляция ответа на внешние стимулы, дифференцировки и биосинтетических процессов. Возможно, результат низкой экспрессии гладкомышечного актина в кластерах 1 и 3 есть следствие угнетения биосинтетических процессов или отрицательной регуляции ответа на внешние стимулы, в данном случае на профибротический матрикс и добавленный в него TGF-beta (Табл. 24).

В 3 кластере карты белок-белковых взаимодействий были построены для ТФ TBX15 и MSC.

Таблица 25. Сигнальный путь для генов ТФ TBX15

KEGG pathways	description	Count in network	strength	False discovery rate
hsa04810	Regulation of actin cytoskeleton	2/209	1,97	0,0389

Среди генов, контролируемых ТФ TBX15 два гена входят в сигнальный путь, регулирующий актиновый цитоскелет, что так же может объяснять разницу в уровне экспрессии гладкомышечного актина между про- и антифибротическими кластерами (Табл. 25).

Таблица 26. Сигнальные пути для генов ТФ MSC

Reactome pathways	description	Count in network	strength	False discovery rate
HSA-9648895	Response of EIF2AK1 (HRI) to heme deficiency	3/15	2.21	0.0028
HSA-9633012	Response of EIF2AK4 (GCN2) to amino acid deficiency	4/99	1.52	0.0077
HSA-8953897	Cellular responses to external stimuli	6/558	0.94	0.0347

Для ТФ MSC особый интерес представляет сигнальный путь, регулирующий ответ на внешние стимулы, который, по-видимому, вовлечен в адаптацию клеток третьего кластера к изменениям микроокружения. Активность данного пути, наряду с выраженным биосинтезом холестерола, может указывать на специализированную функцию этих клеток в

поддержании мембранной стабильности и их способности эффективно реагировать на внешние сигналы (Табл. 26).

3.7 Выделение субпопуляции

Для разделения клеточных субпопуляций с дальнейшим анализом их функциональных свойств с помощью клеточного сортинга необходимо выявить мембранные маркеры этих субпопуляций. Из базы данных UniProt были экспортированы таблицы, содержащие информацию о локализации белков в клетке. Сопоставив список белков с этой базой, было определено 13 белков, являющиеся мембранными. Для сортинга клеток α-SMA- - субпопуляции может быть предложен белок PDGFRa, так как он преимущественно распределен именно в данной субпопуляции. Для сортинга клеток α-SMA+ - субпопуляции МСК из 10 найденных белков были предложены LIMS2, CRIM1 и CDH2, так как они преимущественно распределены именно в данной субпопуляции.

3.8 Экспресс scRNA-seq для определения предрасположенности к развитию фиброза

Быстрый, нацеленный биоинформатический анализ данных секвенирования биологического образца ткани, взятой у пациента, может помочь спрогнозировать поведение соединительной ткани и реакции на различные виды стимулов. В частности механические. Такой подход может быть актуален в эстетической челюстно-лицевой хирургии, где выраженность фибротических процессов может привести к нежелательным для пациента последствиям, связанным с деформацией форм оперируемых областей.

В случае выявления агрессивного характера реакции соединительной ткани возможно применение клеточных продуктов МСК, как фактора, регулирующего патологический процесс ранозаживления с преимущественным образованием соединительной ткани без восстановления

ее функции. В частности, использование внеклеточных везикул МСК, в качестве БМКП может повлиять на дифференцировку фибробластов в миофибробласты, снижая выраженность этого направления. Также, при выявлении достаточно представленной субпопуляции клеток, сдерживающих развитие фибротических изменений, возможна стимуляция этих клеток с целью снижения скорости непродуктивной репарации ткани.

Глава 4. ОБСУЖДЕНИЕ

Мезенхимные стромальные клетки являются ключевыми регуляторами функциональной активности множества тканей организма человека. За счет способности дифференцироваться в адипоциты, остеобласты, хондробласты и секретировать множество факторов роста и цитокинов МСК являются ключевыми участниками регуляции репаративных процессов. При этом основная роль в реализации их регуляторных эффектов отводится биологически активным компонентам секретома МСК, включая пул некодирующих регуляторных РНК, которые будучи перенесенными в составе внеклеточных везикул в другие клетки, способны перепрограммировать их в направлении стимуляции регенеративных процессов. Однако показано, что в ответ на различные сигналы от поврежденных тканей МСК могут сами дифференцироваться в миофибробласты, а также реагировать специфическими изменениями своего секретома. Фиброз органов вносит большой вклад в смертность во всем мире, при этом основную роль в этих процессах играют миофибробласты, которые дифференцируются под действием профибротических сигналов из фибробластов и других типов клеток стромы. Представление о том, что фиброз в разных тканях обусловлен практически идентичной популяцией МСК и схожим набором внешних сигналов постепенно устаревает, и анатомическое распределение и организация резидентных популяций МСК в разных органах определяются во время эмбрионального развития уникальной

комбинацией внешних сигналов, что в конечном итоге формирует совершенно определенную тканеспецифичность.

В рамках данной работы проанализирована гетерогенность ответа мезенхимных стромальных клеток на профибротические стимулы с использованием технологии секвенирования РНК одиночных клеток. Было показано, что в популяции МСК присутствуют субпопуляции, по-разному отвечающие на профибротические сигналы. При культивировании МСК в модели профибротического микроокружения, включающей воздействие децеллюляризованного внеклеточного матрикса дермальных фибробластов человека в присутствии TGFβ, было выявлено шесть различных кластеров клеток. Критически важным открытием стало то, что только часть МСК дифференцировалась в миофибробласты, что определялось по экспрессии специфического маркера α-гладкомышечного актина, в то время как остальные клетки обладали транскриптомным профилем, отличным от характерного для миофибробластов. Это наблюдение фундаментально противоречит традиционному представлению о том, что все МСК в профибротическом микроокружении дифференцируются в миофибробласты. Было показано, что клетки с высоким уровнем экспрессии гена ACTA2 преимущественно располагались во 2 и 4 кластерах, в то время как в 0, 1, 3 и 5 кластерах сосредоточились МСК с низким уровнем экспрессии этого гена. Среднее значение и кратность изменения уровня экспрессии гена ACTA2 в кластерах 2 и 4 в среднем выше более чем в 7 раз по сравнению с остальными кластерами, что позволило четко разделить популяцию на профибротические и антифибротические субпопуляции.

Детальная характеристика субпопуляции МСК, не дифференцирующейся в миофибробласты под действием профибротических стимулов, выявила уникальные транскрипционные особенности этих клеток. Функциональный анализ показал, что данная субпопуляция характеризуется повышенной экспрессией групп генов, отвечающих за организацию и ремоделирование внеклеточного матрикса, регуляцию метаболических

процессов и ангиогенеза. Среди дифференциально экспрессированных генов особое внимание привлекли катепсин К, который участвует в разрушении компонентов внеклеточного матрикса, матриксные металлопептидазы MMP1 и MMP3, обладающие антифибротическим действием благодаря способности расщеплять коллагены различных типов, а также FBLN1, играющий роль в клеточной адгезии и транспорте по волокнам внеклеточного матрикса. Важным наблюдением стала экспрессия длинной некодирующей РНК LINC01705, которая взаимодействует с микроРНК, участвующими в регуляции процессов фиброза. Анализ терминов генной онтологии высокоэкспрессированных генов в α-SMA-негативной субпопуляции показал обогащение процессами дифференцировки клеток, позитивной регуляции клеточной дифференцировки, организации супрамолекулярных волокон и организации внеклеточного матрикса, что соответствует ожидаемому эффекту сдерживания профибротических изменений путем разрушения образующихся в процессе фиброза внеклеточных матриксных белков.

В современной литературе традиционно обсуждается концепция относительно единообразного ответа мезенхимных стромальных клеток на профибротические стимулы. Так, исследования показывают, что TGFβ рассматривается как главный регулятор дифференцировки миофибробластов при фиброзе, что доказывается несколькими исследованиями на различных моделях фиброза. Считается, что популяция миофибробластов гетерогенна и происходит из перицитов, резидентных фибробластов и клеток костного мозга, при этом МСК рассматриваются как один из основных источников миофибробластов при патологическом фиброзе. Однако, полученные данные показывают иные закономерности в поведении МСК в профибротическом микроокружении. Было выявлено, что в ответ на профибротические стимулы популяция МСК не демонстрирует единообразного ответа, а разделяется на дискретные субпопуляции с принципиально различными функциональными характеристиками. Данный механизм гетерогенного ответа сам по себе является не стандартным, так как обычно предполагается, что все клетки

одного типа в одинаковых условиях должны демонстрировать схожий ответ. Результаты расширяют понимание механизма ответа МСК на профибротические стимулы и предполагают существование внутренних регуляторных механизмов, определяющих судьбу отдельных клеток в популяции. Особенно важным является то, что обнаруженная гетерогенность не является случайной, а имеет четкую молекулярную основу, что подтверждается воспроизводимостью результатов при использовании клеток от разных доноров и в различных экспериментальных условиях.

Также был проанализирован сигнальный каскад, активация которого приводит к формированию антифибротической субпопуляции МСК. Показано, что ключевые транскрипционные факторы активируют сигнальные каскады, связанные с организацией внеклеточного матрикса и ангиогенезом, но не с мышечным сокращением и формированием стрессовых волокон, характерных для миофибробластов. Анализ регуляторных сетей с использованием алгоритма SCENIC позволил идентифицировать активные регулоны в различных субпопуляциях МСК. В антифибротической субпопуляции была выявлена активность транскрипционных факторов, участвующих в подавлении фибротического ответа в ответ на повреждение, стимуляции жизнеспособности, пролиферации и дифференцировки клеток. Особое внимание было уделено анализу микроРНК, потенциально вовлеченных в регуляцию дифференцировки и функциональных свойств МСК. Были идентифицированы микроРНК hsa-mir-10a-5p, hsa-mir-34a-5p, hsa-mir-27a-3p, hsa-mir-194-5p, hsa-mir-18a-5p, hsa-mir-20a-5p, hsa-mir-451a, hsa-mir-129-5p, hsa-mir-29a-3p, hsa-mir-29b-3p, hsa-mir-29c-3p, которые, согласно литературным данным, участвуют в регуляции фибротических процессов. Анализ траекторий развития мезенхимальных стволовых клеток с использованием RNA velocity продемонстрировал последовательную активацию клеточных популяций, начиная от клеток, участвующих в организации внеклеточного матрикса и ангиогенезе, и заканчивая клетками, отвечающими за мышечное сокращение и восстановление тканей. Активация

процессов пролиферации и деления клеток в одном из кластеров свидетельствует о важной роли данных клеток в регенерации тканей, подчеркивая их центральное значение в восстановительных процессах.

Большое количество дегенеративных расстройств, связанных с фиброзом, остаются заболеваниями с высокой заболеваемостью и смертностью из-за медленного развития методов лечения этого нарушения. В настоящее время не существует антифибротических препаратов, которые могли бы полностью остановить прогрессирование фибротических изменений, а доступная терапия основана на ингибировании рецепторных тирозинкиназ, которые играют решающую роль в передаче клеточных сигналов в здоровых клетках и тканях. Выявленная субпопуляция МСК с антифибротическими свойствами открывает принципиально новые терапевтические возможности. Быстрый, нацеленный биоинформатический анализ данных секвенирования биологического образца ткани, взятой у пациента, может помочь спрогнозировать поведение соединительной ткани и реакции на различные виды стимулов. При выявлении достаточно представленной субпопуляции клеток, сдерживающих развитие фибротических изменений, возможна стимуляция этих клеток с целью снижения скорости непродуктивной репарации ткани. Новизна и практическая значимость работы подтверждаются полученным патентом РФ №2766707 от 15.03.2022 г. «Средство для лечения фиброза тканей на основе компонентов секретома мезенхимных стромальных клеток, способ получения и применения средства».

Следует подчеркнуть, что участие продемонстрированных механизмов в регуляторных функциях МСК in vivo все еще остается открытым вопросом. Использованная модель профибротического микроокружения, несмотря на включение ключевых компонентов фибротического процесса, не может полностью воспроизвести сложность тканевого микроокружения in vivo, а их влияние, а также взаимодействие с другими типами клеток является предметом дальнейших исследований. Одним из основных ограничений

исследования является использование модели in vitro, которая не может полностью воспроизвести сложность тканевого микроокружения in vivo. Профибротическая модель, основанная на децеллюляризованном внеклеточном матриксе дермальных фибробластов в присутствии TGFβ, хотя и включает ключевые компоненты фибротического процесса, не учитывает влияние иммунных клеток, эндотелиальных клеток, нервных окончаний и других компонентов тканевого микроокружения. Потеря пространственной информации при диссоциации тканей исключает возможность изучения влияния межклеточных взаимодействий и нишевых факторов на поведение МСК. Технические ограничения метода scRNA-seq включают высокий процент нулевых значений, характерный для МСК с их относительно низкой транскрипционной активностью, что может приводить к ложноотрицательным результатам при анализе дифференциальной экспрессии. Batch-эффекты представляют серьезную проблему для исследований МСК, поскольку стромальные клетки крайне чувствительны к условиям культивирования. Биологические ограничения интерпретации связаны с snapshot природой scRNA-seq данных, в то время как дифференцировка МСК представляет собой длительный динамический процесс. Фундаментальный вопрос касается интерпретации клеточной идентичности: неясно, являются ли выявленные кластеры дискретными типами клеток или представляют континуум состояний. Пластичность МСК затрудняет четкую классификацию, а алгоритмы кластеризации навязывают дискретность на потенциально непрерывные данные.

Дальнейшие исследования должны быть направлены на валидацию выявленных механизмов в условиях in vivo с использованием животных моделей фиброза различных органов. Перспективным представляется изучение возможности модуляции соотношения про- и антифибротических субпопуляций МСК с помощью фармакологических агентов или генной терапии. Необходимы дополнительные исследования для определения эпигенетических механизмов, контролирующих судьбу МСК в

профибротическом микроокружении. Особый интерес представляет изучение роли выявленных микроРНК в межклеточной коммуникации через внеклеточные везикулы. Разработка методов селективного выделения и культивирования антифибротической субпопуляции МСК может открыть новые возможности для клеточной терапии фиброза. Интеграция данных scRNA-seq с пространственной транскриптомикой позволит лучше понять роль тканевого микроокружения в определении судьбы МСК. Создание более совершенных моделей профибротического микроокружения, включающих различные типы клеток и факторы, приблизит экспериментальные условия к патофизиологическим процессам in vivo. Необходимо также изучение гетерогенности ответа МСК из различных тканевых источников на профибротические стимулы для определения универсальности выявленных механизмов.

В заключение следует отметить, что полученные результаты фундаментально меняют понимание роли МСК в развитии фиброза. Выявление субпопуляции МСК с антифибротическими свойствами открывает новые перспективы для разработки персонализированных подходов к лечению фибротических заболеваний. Гетерогенность ответов МСК на профибротические стимулы может оказывать критическое влияние на развитие фиброза тканей за счет регуляции баланса между субпопуляциями клеток, пополняющих пул миофибробластов, и клеток с антифибротическими свойствами. Установленные молекулярные маркеры позволяют разделять субпопуляции МСК с использованием клеточного сортинга, что создает основу для разработки новых клеточных технологий. Выявленные регуляторные сети и сигнальные пути представляют потенциальные мишени для фармакологического воздействия. Результаты работы демонстрируют возможности современных одноклеточных технологий для фундаментального изучения клеточных механизмов патологических процессов и подчеркивают необходимость пересмотра традиционных представлений о роли МСК в фиброзе.

ЗАКЛЮЧЕНИЕ

В работе было обнаружено, что клетки образца МСК, культивируемые в профибротических условиях, неоднозначно воспринимают стимулы и проявляется гетерогенность в ответе. Гетерогенность проявляется в возникновении субпопуляции, резко отличающейся по экспрессии специфического для миофибробластов маркера – aSMA (гладкомышечного актина). Так как миофибробласты являются основными клетками, опосредующими развитие фиброза, было предположено, что клетки, в которых понижена экспрессия гладкомышечного актина, возможно, наделяются антифибротическими свойствами.

Для дальнейшего детального изучения необходимо типировать клетки образца, проанализировать биологические процессы, происходящие в клетках.

Появление методов транскриптомики единичных клеток открыло новые возможности в анализе данных секвенирования. Возможность анализа профиля экспрессии отдельных клеток позволило обнаружить в исследуемых образцах не обнаруживаемые ранее субпопуляции клеток. Одновременно с появившимися возможностями и преимуществами новых методов наука столкнулась с новыми проблемами и задачами.

Каждая клетка в образце находится в своем индивидуальном состоянии и это состояние совершенно не обязательно должно быть начальным, из которого она начинает свой путь дифференцировки или конечным, в который она стремится. Даже если большое количество отдельных клеток стремятся дифференцироваться в определенный тип клеток, каждая из них в каждый момент времени будет находится в индивидуальном состоянии.

В стандартном анализе картирование последовательностей транскриптов происходит на кодирующие участки генома, то есть не учитываются транскрипты, выравниваемые с интронными и участками и участками, находящимися между интронами и экзонами. В таком режиме отсутствует

возможность учитывать транскрипционную динамику и, следовательно нельзя прогнозировать будущее состояние клеток.

При анализе данных секвенирования единичных клеток включенные в алгоритм обработки данных определяют уровень экспрессии каждого транскрипта в каждой клетке и затем, ориентируясь на меру сходства между уровнями экспрессий отдельных клеток между разными клетками объединяет их в схожие группы, которые называются кластерами. В результате работы ожидается, что в одном кластере окажутся клетки одного типа. В действительности в кластере находится лишь небольшая часть клеток в конечном дифференцированном состоянии. По этой причине нельзя с уверенностью типировать клетки в каждом из кластеров.

В описываем алгоритме сначала клетки объединяются в кластеры по мере схожести паттернов экспрессий. Затем определяются высокоэкспрессированные гены в каждом кластере. Определяемые гены должны помочь нам идентифицировать клетки как относящиеся к какому-то из типов клеток. На практике это реализуется с трудом по описанной выше причине – лишь небольшая часть клеток в кластере находятся в конечно дифференцированном состоянии. Для того, чтобы выйти из сложившейся ситуации можно использовать кластеризацию высокопредставленных генов кластера по биологическим процессам. Информация о протекающих в клетке биологических процессах помогает понять, что в клетке происходит в данный момент и, возможно, позволяет предположить для чего в ней запущены определенные процессы.

Следовательно, нельзя рассчитывать на точность определения типа клетки используя для идентификации специфические маркеры и актуальные биологические процессы.

Предложенная методика использования данных картирования транскриптов не только на экзонные части, но также на интронные и промежуточные позволило моделировать транскрипционную динамику, основываясь на знаниях о том, что каждая мРНК претерпевает в клетке

процессинг, во время которого она сначала транскрибируется с исходной последовательности ДНК, а затем превращается в свою зрелую форму. Метод транскрипционной динамики или RNA velocity учитывает соотношений нейсплайсированных и сплайсированных форм РНК. Таким образом, состояние индукции гена можно описать как преобладание транскрипции над сплайсингом и деградацией, а состоянии репрессии гена, как преобладание деградации РНК над транскрипцией. В результате такого подхода можно расположить все клетки образца вдоль некоторой линии, которая будет строится из последовательных состояний клеток от исходного состояния к конечному. Такие линии называются траекториями развития. Анализ траекторий развития возможно в латентном времени и псевдовремени. Анализ будет отличаться тем, что в латентном времени выявляются временные отличия между различными судьбами развития клеток, а в псевдовремени – только распределение клеток во временном континууме, представляющего собой отрезок времени от начального состояния клетки до конечно дифференцированного.

Интересной особенностью использования современных методик анализа данных транскриптома единичных клеток является возможность получения списка переходных генов, которые определяются алгоритмом как наиболее представленные в конкретные моменты времени, называемые точками ветвления. Такой подход концептуально схож с типированием клеток по биологическим процессам, с той лишь разницей, что при типировании по биологическим процессам получаемый список генов никак не связан с критическими точками в развитии клетки, а при получении списка переходных генов – связан, что дает нам более точное представление о том, какие конкретно гены ответственны за переход клетки из одного состояния в другое.

В работе частично удалось типировать клетки по специфическим маркерам. Хорошо типировались кластеры, в которых были высоко представлены гены, используемые для идентификации типов клеток. В

остальных случаях такую информацию получить не удалось, так как высокоэкспрессированные гены в других кластерах не являлись специфическими маркерами клеток, а отражали текущие изменения метаболизма клетки.

Изучение актуальных биологических процессов в клетке дополнило информацию. Полученную после типирования по специфическим маркерам. Однако в результатах остались неопознанные кластеры, часть из которых удалось идентифицировать применив еще один метод типии рования клеток, а именно типирование кластеров, находящихся на пути траекторий развития. В таком случае, если тип клеток кластера, находящегося между кластерами с известными типами, неизвестен, можно предположить, что он является частью дифференцировочного пути клеток и определить тип клеток, составляющих его.

Таким образом, несмотря на большие возможности новых методов изучения единичных клеток и анализа данных секвенирования, нельзя сказать, что ответы на вопросы, имеющиеся до появления этих методов найдены. По-прежнему остается неоднозначной задача типирования клеток. Возможно, основным ограничивающим фактором на пути решения этого вопроса является изначально неправильная методология определения типов клеток, основанная на специфических маркерах. Скорее всего идентификация типов клеток — это системный вопрос, который требует многостороннего подхода. Здесь нужно иметь информацию и о специфических маркерах, и о биологических процессах, о точках ветвления траекторий развития, о переходных генах, характеризующих важные изменения транскрипционного профиля клетки, открытость и закрытость участков хроматина, присутствие транскриптов транскрипционных факторов и регуляторных элементов.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1) С помощью метода scRNA-seq выявлены субпопуляции МСК, дифференцирующиеся и не дифференцирующиеся в миофибробласты при культивировании в модели профибротического микроокружения.

2) Анализ транскриптомного паттерна субпопуляций МСК позволил типировать в выявленных кластерах миофибробласты, хондроциты, эндотелиальные клетки, фибробласты, миоциты и гладкомышечные клетки, а также предположить, что субпопуляция МСК, которая не дифференцируется в миофибробласты в профибротических условиях, преимущественно вовлечена в процессы, связанные с организацией и ремоделированием ВКМ, регуляцией ангиогенеза, обмена липидов, дифференцировкой клеток в остеогенном и хондрогенном направлениях. При этом по результатам анализа регулонов в клетках этой субпопуляции была установлена активность ряда транскрипционных факторов, участвующих в подавлении фибротического ответа в ответ на повреждение, стимуляции жизнеспособности, пролиферации и дифференцировки клеток.

3) Анализ траекторий развития МСК, культивируемых в профибротических условиях, подтвердил гетерогенность ответа клеток на сигналы от микроокружения в виде разделения на различные по транскриптомному паттерну клеточные субпопуляции стромальных клеток.

4) Анализ экспрессии нкРНК в субпопуляциях МСК, по-разному отвечающих на профибротические сигналы, позволил выявить днРНК (LINC01705) и ряд микроРНК (hsa-mir-10a-5p, hsa-mir-34a-5p, hsa-mir-27a-3p, hsa-mir-194-5p, hsa-mir-18a-5p, hsa-mir-20a-5p, hsa-mir-451a, hsa-mir-129-5p, hsa-mir-29a-3p, hsa-mir-29b-3p, hsa-mir-29c-3p), потенциально вовлеченные в регуляцию дифференцировки и функциональных свойств МСК, не дифференцирующихся в миофибробласты.

5) Были установлены мембранные белки (PDGFRa, LIMS2, CRIM1 и CDH2), экспрессия которых значимо различается в субпопуляциях МСК, по-разному отвечающих на профибротические сигналы, что позволяет

предложить их в качестве маркеров для разделения субпопуляций с использованием клеточного сортинга.

СПИСОК СОКРАЩЕНИЙ

ВВ – внеклеточные везикулы

ВКМ – внеклеточный матрикс

МСК – мезенхимные стромальные клетки

КОЕ-Ф – колониеобразующая единица фибробластов

CD – кластер дифференцировки

Rs-клетки - rapidly self-renewing cells

scRNA-seq - Single-cell RNA sequencing

UMI - Unique molecular identifiers

FACS - Fluorescence-activated cell sorting

DMEM - Gibco Dulbecco's Modified Eagle Medium

GEM - Gel Bead-in-Emulsion

PCA - Principal component analysis

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Owen M., Friedenstein A. J. Stromal stem cells: marrow-derived osteogenic precursors //Ciba Found Symp. – 1988. – Т. 136. – №. 29. – С. 42-60.
2. Dominici M., Le Blanc K., Mueller I., Slaper-Cortenbach I., Marini F., Krause D. et al. Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement //Cytotherapy. – 2006. – Т. 8. – №. 4. – С. 315-317.
3. da Silva Meirelles L., Chagastelles P. C., Nardi N. B. Mesenchymal stem cells reside in virtually all post-natal organs and tissues //Journal of cell science. – 2006. – Т. 119. – №. 11. – С. 2204-2213.
4. Crisan M. et al. A perivascular origin for mesenchymal stem cells in multiple human organs //Cell stem cell. – 2008. – Т. 3. – №. 3. – С. 301-313.

5. Haghverdi L. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors //Nature biotechnology. – 2018. – Т. 36. – №. 5. – С. 421-427.
6. Robinson M. D., Smyth G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data //Biostatistics. – 2008. – Т. 9. – №. 2. – С. 321-332.
7. Haghverdi L. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors //Nature biotechnology. – 2018. – Т. 36. – №. 5. – С. 421-427.
8. Park J. E. et al. Fast batch alignment of single cell transcriptomes unifies multiple mouse cell atlases into an integrated landscape //BioRxiv. – 2018. – С. 397042.
9. Robinson M. D., Smyth G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data //Biostatistics. – 2008. – Т. 9. – №. 2. – С. 321-332.
10. Murray I. R., West C. C., Hardy W. R., James A. W., Park T. S., Nguyen A. et al. Natural history of mesenchymal stem cells, from vessel walls to culture vessels //Cell Mol Life Sci. – 2014. – Т. 71. – №. 8. – С. 1353-1374.
11. Caplan A. I. Mesenchymal stem cells //Journal of orthopaedic research. – 1991. – Т. 9. – №. 5. – С. 641-650.
12. Horwitz E. M. et al. Clarification of the nomenclature for MSC: The International Society for Cellular Therapy position statement //Cytotherapy. – 2005. – Т. 7. – №. 5. – С. 393-395.
13. Owen M. E., Cave J., Joyner C. J. Clonal analysis in vitro of osteogenic differentiation of marrow CFU-F //Journal of cell science. – 1987. – Т. 87. – №. 5. – С. 731-738.
14. Lebedinskaia O. V. et al. Analysis of the changes of stromal precursor cell numbers in the thymus and the spleen of animals of different age groups //Morfologiia (Saint Petersburg, Russia). – 2005. – Т. 127. – №. 3. – С. 41-44.

15. Kolesnikova A. I. et al. Radiosensitivity of cells-precursors of hemopoietic stroma (CFU-F) in the rat bone marrow under the effect of 60Co gamma irradiation in various conditions //Radiobiologiia. – 1992. – Т. 32. – №. 6. – С. 844-850.
16. Kuçi Z. et al. Clonal analysis of multipotent stromal cells derived from CD271+ bone marrow mononuclear cells: functional heterogeneity and different mechanisms of allosuppression //Haematologica. – 2013. – Т. 98. – №. 10. – С. 1609-1617.
17. Colter D. C., Sekiya I., Prockop D. J. Identification of a subpopulation of rapidly self-renewing and multipotential adult stem cells in colonies of human marrow stromal cells //Proceedings of the National Academy of Sciences. – 2001. – Т. 98. – №. 14. – С. 7841-7845.
18. Prockop D. J., Sekiya I., Colter D. C. Isolation and characterization of rapidly self-renewing stem cells from cultures of human marrow stromal cells //Cytotherapy. – 2001. – Т. 3. – №. 5. – С. 393-396.
19. Izadpanah R. et al. Characterization of multipotent mesenchymal stem cells from the bone marrow of rhesus macaques //Stem cells and development. – 2005. – Т. 14. – №. 4. – С. 440-451.
20. Neuhuber B. et al. Effects of plating density and culture time on bone marrow stromal cell characteristics //Experimental hematology. – 2008. – Т. 36. – №. 9. – С. 1176-1185.
21. Prockop D. J., Sekiya I., Colter D. C. Isolation and characterization of rapidly self-renewing stem cells from cultures of human marrow stromal cells //Cytotherapy. – 2001. – Т. 3. – №. 5. – С. 393-396.
22. Boxall S. A., Jones E. Markers for characterization of bone marrow multipotential stromal cells //Stem cells international. – 2012. – Т. 2012. – Article ID 975871.
23. Battula V. L. et al. Isolation of functionally distinct mesenchymal stem cell subsets using antibodies against CD56, CD271, and mesenchymal stem cell antigen-1 //Haematologica. – 2009. – Т. 94. – №. 2. – С. 173-184.

24. Russell K. C. et al. In vitro high-capacity assay to quantify the clonal heterogeneity in trilineage potential of mesenchymal stem cells reveals a complex hierarchy of lineage commitment //Stem cells. – 2010. – Т. 28. – №. 4. – С. 788-798.
25. Owen M. E., Cave J., Joyner C. J. Clonal analysis in vitro of osteogenic differentiation of marrow CFU-F //Journal of cell science. – 1987. – Т. 87. – №. 5. – С. 731-738.
26. Phinney D. G. et al. Plastic adherent stromal cells from the bone marrow of commonly used strains of inbred mice: variations in yield, growth, and differentiation //Journal of cellular biochemistry. – 1999. – Т. 72. – №. 4. – С. 570-585.
27. Van den Heuvel R. et al. Stromal cells from murine developing hemopoietic organs: comparison of colony-forming unit of fibroblasts and long-term cultures //International Journal of Developmental Biology. – 2003. – Т. 35. – №. 1. – С. 33-41.
28. Kim Y. H. et al. Characterization of different subpopulations from bone marrow-derived mesenchymal stromal cells by alkaline phosphatase expression //Stem cells and development. – 2012. – Т. 21. – №. 16. – С. 2958-2968.
29. Nifontova I., Svinareva D., Petrova T., Drize N. Sensitivity of mesenchymal stem cells and their progeny to medicines used for the treatment of hematoproliferative diseases //Cellular Therapy and Transplantation. – 2008. – Т. 1. – №. 2. – С. 54-59.
30. Starostin V. I. et al. The 5-fluorouracil sensitivity of the hematopoietic stroma of the bone marrow and spleen //Izvestiia Akademii nauk. Seriia biologicheskaia. – 1995. – №. 4. – С. 496-500.
31. Domaratskaia E. I. et al. Alkylating damage by dipin of hematopoietic and stromal cells of the bone marrow //Izvestiia Akademii nauk. Seriia biologicheskaia. – 2005. – №. 3. – С. 267-272.

32. Ben-Ishay Z. et al. Pre-CFU-f: Young-type stromal stem cells in murine bone marrow following administration of DNA inhibitors //The International Journal of Cell Cloning. – 1986. – Т. 4. – №. 2. – С. 126-134.
33. Conget P. A., Allers C., Minguell J. J. Identification of a discrete population of human bone marrow-derived mesenchymal cells exhibiting properties of uncommitted progenitors //Journal of hematotherapy & stem cell research. – 2001. – Т. 10. – №. 6. – С. 749-758.
34. Domaratskaia E. I. et al. Alkylating damage by dipin of hematopoietic and stromal cells of the bone marrow //Izvestiia Akademii nauk. Seriia biologicheskaia. – 2005. – №. 3. – С. 267-272.
35. Wan C. et al. Nonadherent cell population of human marrow culture is a complementary source of mesenchymal stem cells (MSCs) //Journal of Orthopaedic Research. – 2006. – Т. 24. – №. 1. – С. 21-28.
36. Byeverova E. I., Bragina E. V., Molchanova E. A. Nonadhesive populations in cultures of mesenchymal stromal cells from hematopoietic organs in mouse and rat //Ontogenez. – 2008. – Т. 39. – №. 6. – С. 420-429.
37. MacArthur B. D. et al. A non-invasive method for in situ quantification of subpopulation behaviour in mixed cell culture //Journal of the Royal Society Interface. – 2006. – Т. 3. – №. 6. – С. 63-69.
38. Tanaka-Douzono M. et al. Detection of murine adult bone marrow stroma-initiating cells in Lin− c-fms+ c-kitlowVCAM-1+ cells //Journal of cellular physiology. – 2001. – Т. 189. – №. 1. – С. 45-53.
39. Wan C. et al. Nonadherent cell population of human marrow culture is a complementary source of mesenchymal stem cells (MSCs) //Journal of Orthopaedic Research. – 2006. – Т. 24. – №. 1. – С. 21-28.
40. Molchanova E. A. et al. The sensitivity of mesenchymal stromal cells subpopulations with different adhesion properties and derived from hemopoietic organs to growth factors EGF, bFGF, and PDGF //Biology Bulletin. – 2011. – Т. 38. – №. 2. – С. 99-108.

41. Owen M., Friedenstein A. J. Stromal stem cells: marrow-derived osteogenic precursors //Ciba Found Symp. – 1988. – T. 136. – №. 29. – C. 42-60.
42. Muraglia A., Cancedda R., Quarto R. Clonal mesenchymal progenitors from human bone marrow differentiate in vitro according to a hierarchical model //Journal of cell science. – 2000. – T. 113. – №. 7. – C. 1161-1166.
43. De Bari C. et al. Multipotent mesenchymal stem cells from adult human synovial membrane //Arthritis & Rheumatism. – 2001. – T. 44. – №. 8. – C. 1928-1942.
44. Russell K. C. et al. In vitro high-capacity assay to quantify the clonal heterogeneity in trilineage potential of mesenchymal stem cells reveals a complex hierarchy of lineage commitment //Stem cells. – 2010. – T. 28. – №. 4. – C. 788-798.
45. DiGirolamo C. M. et al. Propagation and senescence of human marrow stromal cells in culture: a simple colony-forming assay identifies samples with the greatest potential to propagate and differentiate //British journal of haematology. – 1999. – T. 107. – №. 2. – C. 275-281.
46. Sekiya I. et al. Expansion of human adult stem cells from bone marrow stroma: conditions that maximize the yields of early progenitors and evaluate their quality //Stem cells. – 2002. – T. 20. – №. 6. – C. 530-541.
47. Gronthos S. et al. Molecular and cellular characterisation of highly purified stromal stem cells derived from human bone marrow //Journal of cell science. – 2003. – T. 116. – №. 9. – C. 1827-1835.
48. DiGirolamo C. M. et al. Propagation and senescence of human marrow stromal cells in culture: a simple colony-forming assay identifies samples with the greatest potential to propagate and differentiate //British journal of haematology. – 1999. – T. 107. – №. 2. – C. 275-281.
49. Gregory C. A., Ylostalo J., Prockop D. J. Adult bone marrow stem/progenitor cells (MSCs) are preconditioned by microenvironmental "niches" in culture: a two-stage hypothesis for regulation of MSC fate //Science Signaling. – 2005. – T. 2005. – №. 294. – C. pe37.

50. Friedman S. L. et al. Therapy for fibrotic diseases: nearing the starting line //Science translational medicine. – 2013. – Т. 5. – №. 167. – С. 167sr1.
51. Baglama J., Reichel L. Augmented implicitly restarted Lanczos bidiagonalization methods //SIAM Journal on Scientific Computing. – 2005. – Т. 27. – №. 1. – С. 19-42.
52. Van Der Maaten L. Accelerating t-SNE using tree-based algorithms //The Journal of Machine Learning Research. – 2014. – Т. 15. – №. 1. – С. 3221-3245.
53. McInnes L., Healy J., Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction //arXiv preprint arXiv:1802.03426. – 2018.
54. Blondel V. D. et al. Fast unfolding of communities in large networks //Journal of statistical mechanics: theory and experiment. – 2008. – Т. 2008. – №. 10. – С. P10008.
55. Yu D., Huber W., Vitek O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size //Bioinformatics. – 2013. – Т. 29. – №. 10. – С. 1275-1282.
56. Hinz B. et al. The myofibroblast: one function, multiple origins //The American journal of pathology. – 2007. – Т. 170. – №. 6. – С. 1807-1816.
57. Humphreys B. D. et al. Fate tracing reveals the pericyte and not epithelial origin of myofibroblasts in kidney fibrosis //The American journal of pathology. – 2010. – Т. 176. – №. 1. – С. 85-97.
58. Moore-Morris T. et al. Resident fibroblast lineages mediate pressure overload–induced cardiac fibrosis //The Journal of clinical investigation. – 2014. – Т. 124. – №. 7. – С. 2921-2934.
59. Rock J. R. et al. Multiple stromal populations contribute to pulmonary fibrosis without evidence for epithelial to mesenchymal transition //Proceedings of the National Academy of Sciences. – 2011. – Т. 108. – №. 52. – С. E1475-E1483.
60. Grupp C. et al. A novel model to study renal myofibroblast formation in vitro //Kidney international. – 2001. – Т. 59. – №. 2. – С. 543-553.
61. Hoyles R. K. et al. An essential role for resident fibroblasts in experimental lung fibrosis is defined by lineage-specific deletion of high-affinity type II

transforming growth factor β receptor //American journal of respiratory and critical care medicine. – 2011. – Т. 183. – №. 2. – С. 249-261.

62. Kanzler S. et al. TGF-β1 in liver fibrosis: an inducible transgenic mouse model to study liver fibrogenesis //American Journal of Physiology-Gastrointestinal and Liver Physiology. – 1999. – Т. 276. – №. 4. – С. G1059-G1068.
63. Kim K. K. et al. Alveolar epithelial cell mesenchymal transition develops in vivo during pulmonary fibrosis and is regulated by the extracellular matrix //Proceedings of the National Academy of Sciences. – 2006. – Т. 103. – №. 35. – С. 13180-13185.
64. Lijnen P. J., Petrov V. V., Fagard R. H. Induction of cardiac fibrosis by transforming growth factor-β1 //Molecular genetics and metabolism. – 2000. – Т. 71. – №. 1-2. – С. 418-435.
65. El Agha E. et al. Two-way conversion between lipogenic and myogenic fibroblastic phenotypes marks the progression and resolution of lung fibrosis //Cell stem cell. – 2017. – Т. 20. – №. 2. – С. 261-273.e3.
66. Friedenstein A. J., Piatetzky-Shapiro I. I., Petrakova K. V. Osteogenesis in transplants of bone marrow cells //Journal of Embryology and Experimental Morphology. – 1966. – Т. 16. – №. 3. – С. 381-390.
67. Raposo G., Stoorvogel W. Extracellular vesicles: exosomes, microvesicles, and friends //Journal of Cell Biology. – 2013. – Т. 200. – №. 4. – С. 373-383.
68. Raposo G. et al. B lymphocytes secrete antigen-presenting vesicles //Journal of Experimental Medicine. – 1996. – Т. 183. – №. 3. – С. 1161-1172.
69. King Jr T. E. et al. A phase 3 trial of pirfenidone in patients with idiopathic pulmonary fibrosis //New England Journal of Medicine. – 2014. – Т. 370. – №. 22. – С. 2083-2092.
70. van Manen M. J. G. et al. Effect of nintedanib on pulmonary function in patients with idiopathic pulmonary fibrosis and severe physiologic impairment //Respiratory Medicine. – 2017. – Т. 131. – С. 91-98.

71. Richeldi L. et al. Efficacy and safety of nintedanib in idiopathic pulmonary fibrosis //New England Journal of Medicine. – 2014. – Т. 370. – №. 22. – С. 2071-2082.
72. Nelson D. L., Cox M. M. Lehninger Principles of Biochemistry. – New York: W. H. Freeman, 2013. – 1328 p.
73. García-López J., Brieño-Enríquez M. A., Del Mazo J. MicroRNA biogenesis and variability //Biomolecular Concepts. – 2013. – Т. 4. – №. 4. – С. 367-380.
74. Ambros V. et al. A uniform system for microRNA annotation //RNA. – 2003. – Т. 9. – №. 3. – С. 277-279.
75. Almeida M. I., Reis R. M., Calin G. A. MicroRNA history: discovery, recent applications, and next frontiers //Mutation Research. – 2011. – Т. 717. – №. 1-2. – С. 1-8.
76. Thomas J. MicroRNA Nomenclature //News-Medical. – 2010. – URL: https://www.news-medical.net/life-sciences/MicroRNA-Nomenclature.aspx (дата обращения: 05.05.2021).
77. García-López J., Brieño-Enríquez M. A., Del Mazo J. MicroRNA biogenesis and variability //Biomolecular Concepts. – 2013. – Т. 4. – №. 4. – С. 367-380.
78. García-López J., Brieño-Enríquez M. A., Del Mazo J. MicroRNA biogenesis and variability //Biomolecular Concepts. – 2013. – Т. 4. – №. 4. – С. 367-380.
79. García-López J., Brieño-Enríquez M. A., Del Mazo J. MicroRNA biogenesis and variability //Biomolecular Concepts. – 2013. – Т. 4. – №. 4. – С. 367-380.
80. Havens M. A., Reich A. A., Duelli D. M., Hastings M. L. Biogenesis of mammalian microRNAs by a non-canonical processing pathway //Nucleic Acids Research. – 2012. – Т. 40. – №. 10. – С. 4626-4640.
81. Havens M. A., Reich A. A., Duelli D. M., Hastings M. L. Biogenesis of mammalian microRNAs by a non-canonical processing pathway //Nucleic Acids Research. – 2012. – Т. 40. – №. 10. – С. 4626-4640.
82. García-López J., Brieño-Enríquez M. A., Del Mazo J. MicroRNA biogenesis and variability //Biomolecular Concepts. – 2013. – Т. 4. – №. 4. – С. 367-380.

83. Schwarz D. S., Hutvágner G., Du T., Xu Z., Aronin N., Zamore P. D. Asymmetry in the assembly of the RNAi enzyme complex //Cell. – 2003. – T. 115. – №. 2. – C. 199-208.
84. Lin S.-L., Chang D., Ying S.-Y. Asymmetry of intronic pre-miRNA structures in functional RISC assembly //Gene. – 2005. – T. 356. – C. 32-38.
85. Fehlmann T. et al. Common diseases alter the physiological age-related blood microRNA profile //Nature Communications. – 2020. – T. 11. – №. 1. – Article 5958.
86. García-López J., Brieño-Enríquez M. A., Del Mazo J. MicroRNA biogenesis and variability //Biomolecular Concepts. – 2013. – T. 4. – №. 4. – C. 367-380.
87. Morin R. D. et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells //Genome Research. – 2008. – T. 18. – №. 4. – C. 610-621.
88. García-López J., Brieño-Enríquez M. A., Del Mazo J. MicroRNA biogenesis and variability //Biomolecular Concepts. – 2013. – T. 4. – №. 4. – C. 367-380.
89. Fischer S. E. J. RNA Interference and MicroRNA-Mediated Silencing //Current Protocols in Molecular Biology. – 2015. – T. 112. – C. 26.1.1-26.1.5.
90. Mohr A. M., Mott J. L. Overview of microRNA biology //Seminars in Liver Disease. – 2015. – T. 35. – №. 1. – C. 3-11.
91. Shekhar K., Menon V. Identification of cell types from single-cell transcriptomic data //Computational Methods for Single-Cell Data Analysis. – New York: Humana Press, 2019. – C. 45-77.
92. Ma F., Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing //Bioinformatics. – 2020. – T. 36. – №. 2. – C. 533-538.
93. Dong X. et al. Cell Type Identification from Single-Cell Transcriptomic Data via Semi-supervised Learning //arXiv preprint arXiv:2005.03994. – 2020.
94. Trapnell C. Defining cell types and states with single-cell genomics //Genome research. – 2015. – T. 25. – №. 10. – C. 1491-1498.
95. Peyvandipour A. et al. Identification of cell types from single cell data using stable clustering //Scientific reports. – 2020. – T. 10. – №. 1. – Article 12023.

96. Pettit J. B. et al. Identifying cell types from spatially referenced single-cell expression datasets //PLoS Computational Biology. – 2014. – T. 10. – №. 9. – Article e1003824.
97. Miao Z. et al. Putative cell type discovery from single-cell gene expression data //Nature Methods. – 2020. – T. 17. – №. 6. – C. 621-628.
98. Zheng G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells //Nature Communications. – 2017. – T. 8. – Article 14049.
99. Wolf F. A., Angerer P., Theis F. J. SCANPY: large-scale single-cell gene expression data analysis //Genome Biology. – 2018. – T. 19. – №. 1. – Article 15.
100. McCarthy D. J., Campbell K. R., Lun A. T., Wills Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R //Bioinformatics. – 2017. – T. 33. – №. 8. – C. 1179-1186.
101. Hao Y. et al. Integrated analysis of multimodal single-cell data //Cell. – 2021. – T. 184. – №. 13. – C. 3573-3587.e29.
102. Satija R. et al. Spatial reconstruction of single-cell gene expression data //Nature biotechnology. – 2015. – T. 33. – №. 5. – C. 495-502.
103. Butler A. et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species //Nature biotechnology. – 2018. – T. 36. – №. 5. – C. 411-420.
104. Stuart T. et al. Comprehensive integration of single-cell data //Cell. – 2019. – T. 177. – №. 7. – C. 1888-1902.e21.
105. Hao Y. et al. Integrated analysis of multimodal single-cell data //bioRxiv. – 2020. – doi: 10.1101/2020.10.12.335331.
106. Pereira W. J. et al. Asc-Seurat: analytical single-cell Seurat-based web application //BMC Bioinformatics. – 2021. – T. 22. – Article 556.
107. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update //Nucleic Acids Research. – 2022. – T. 50. – №. W1. – C. W345-W351.

108. Barkas N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections //Nature Methods. – 2019. – Т. 16. – №. 8. – С. 695-698.
109. Barkas N., Petukhov V., Nikolaeva D., Kharchenko P., Biederstedt E. pagoda2: Single Cell Analysis and Differential Expression. R package version 1.0.10. – 2021.
110. Petukhov V. et al. dropEst: pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments //Genome Biology. – 2018. – Т. 19. – Article 78.
111. La Manno G. et al. RNA velocity of single cells //Nature. – 2018. – Т. 560. – №. 7719. – С. 494-498.
112. Bergen V., Soldatov R. A., Kharchenko P. V., Theis F. J. RNA velocity—current challenges and future perspectives //Molecular Systems Biology. – 2021. – Т. 17. – №. 8. – Article e10282.
113. Bergen V. et al. Generalizing RNA velocity to transient cell states through dynamical modeling //Nature Biotechnology. – 2020. – Т. 38. – №. 12. – С. 1408-1414.
114. Tian T. et al. Clustering single-cell RNA-seq data with a model-based deep learning approach //Nature Machine Intelligence. – 2019. – Т. 1. – №. 4. – С. 191-198.
115. Wang L. et al. scCapsNet-mask: an updated version of scCapsNet with extended applicability in functional analysis related to scRNA-seq data //Research Square. – 2022. – doi: 10.21203/rs.3.rs-1763879/v1.
116. Zheng G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells //Nature Communications. – 2017. – Т. 8. – Article 14049.
117. Illumina. bcl2fastq2 Conversion Software v2.20 Software Guide (Document #15051736). – San Diego: Illumina, Inc., 2018.
118. Haghverdi L. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors //Nature biotechnology. – 2018. – Т. 36. – №. 5. – С. 421-427.

119. Park J. E. et al. Fast batch alignment of single cell transcriptomes unifies multiple mouse cell atlases into an integrated landscape //BioRxiv. – 2018. – doi: 10.1101/397042.
120. Hie B., Bryson B., Berger B. Panoramic stitching of heterogeneous single-cell transcriptomic data //BioRxiv. – 2018. – doi: 10.1101/371179.
121. Haghverdi L. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors //Nature biotechnology. – 2018. – Т. 36. – №. 5. – С. 421-427.
122. Baglama J., Reichel L. Augmented implicitly restarted Lanczos bidiagonalization methods //SIAM Journal on Scientific Computing. – 2005. – Т. 27. – №. 1. – С. 19-42.
123. Scholz M. Approaches to analyse and interpret biological profile data: Dissertation. – Potsdam: Universität Potsdam, 2006. – 157 p.
124. Fang R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC //Nature Communications. – 2021. – Т. 12. – Article 1337.
125. Sun S., Zhu J., Ma Y., Zhou X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis //Genome Biology. – 2019. – Т. 20. – Article 269.
126. Ding J., Condon A., Shah S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models //Nature Communications. – 2018. – Т. 9. – Article 2002.
127. Arbatsky M. et al. Points of Significance: Principal Component Analysis for Biocentric Data Visualization //BioNanoScience. – 2022. – Т. 12. – №. 4. – С. 1366-1380.
128. Yao F., Coquery J., Lê Cao K. A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets //BMC Bioinformatics. – 2012. – Т. 13. – Article 24.
129. Huang H. et al. Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization //Communications Biology. – 2022. – Т. 5. – Article 719.

130. Van der Maaten L., Hinton G. Visualizing data using t-SNE //Journal of Machine Learning Research. – 2008. – Т. 9. – №. 11. – С. 2579-2605.
131. McInnes L., Healy J., Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction //arXiv preprint arXiv:1802.03426. – 2018.
132. Simmons S. et al. Discovering What Dimensionality Reduction Really Tells Us About RNA-Seq Data //Journal of Computational Biology. – 2015. – Т. 22. – №. 8. – С. 715-728.
133. Boileau P., Hejazi N. S., Dudoit S. Exploring high-dimensional biological data with sparse contrastive principal component analysis //Bioinformatics. – 2020. – Т. 36. – №. 11. – С. 3422-3430.
134. Lehrmann A. et al. Visualizing dimensionality reduction of systems biology data //Data Mining and Knowledge Discovery. – 2013. – Т. 27. – №. 1. – С. 146-165.
135. Eckmann J. P., Tlusty T. Dimensional reduction in complex living systems: where, why, and how //BioEssays. – 2021. – Т. 43. – №. 9. – Article 2100062.
136. Wikimedia Commons. Hierarchical clustering diagram. – URL: https://commons.wikimedia.org/wiki/File:Hierarchical_clustering_diagram.png (дата обращения: 15.03.2023).
137. Wikipedia. K-means clustering. – URL: https://en.wikipedia.org/wiki/K-means_clustering (дата обращения: 15.03.2023).
138. Kumar P., Chand N. Clustering in wireless multimedia sensor networks //Journal of Sensor Technology. – 2013. – Т. 3. – №. 2. – С. 41-47.
139. Arkhangel'Skii A. et al. General Topology I: Basic Concepts and Constructions Dimension Theory. – London: Springer, 2011. – 202 p.
140. Deza E., Deza M. M. Encyclopedia of Distances. – Berlin: Springer, 2009. – 590 p.
141. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. – New York: Springer, 2001. – 533 p.
142. Ester M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise //Proceedings of the Second International Conference on

Knowledge Discovery and Data Mining (KDD-96). – AAAI Press, 1996. – С. 226-231.

143. Estivill-Castro V. Why so many clustering algorithms: a position paper //ACM SIGKDD Explorations Newsletter. – 2002. – Т. 4. – №. 1. – С. 65-75.
144. Blondel V. D. et al. Fast unfolding of communities in large networks //Journal of Statistical Mechanics: Theory and Experiment. – 2008. – Т. 2008. – №. 10. – Article P10008.
145. MacQueen J. Some methods for classification and analysis of multivariate observations //Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. – 1967. – Т. 1. – №. 14. – С. 281-297.
146. Садовский М. Г., Чернышова А. И. Выявление связи структуры и таксономии геномов хлоропластов методом динамических ядер //Фундаментальные исследования. – 2014. – №. 11-3. – С. 545-549.
147. Chen X., Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method //Bioinformatics. – 2015. – Т. 31. – №. 12. – С. 1974-1980.
148. Levine J. H. et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis //Cell. – 2015. – Т. 162. – №. 1. – С. 184-197.
149. Arya S. et al. An optimal algorithm for approximate nearest neighbor searching //Journal of the ACM. – 1998. – Т. 45. – №. 6. – С. 891-923.
150. Blondel V. D. et al. Fast unfolding of communities in large networks //Journal of Statistical Mechanics: Theory and Experiment. – 2008. – Т. 2008. – №. 10. – Article P10008.
151. Waltman L., Van Eck N. J. A smart local moving algorithm for large-scale modularity-based community detection //The European Physical Journal B. – 2013. – Т. 86. – №. 11. – Article 471.
152. Traag V. A., Waltman L., Van Eck N. J. From Louvain to Leiden: guaranteeing well-connected communities //Scientific Reports. – 2019. – Т. 9. – Article 5233.

153. Flores M. et al. Deep learning tackles single-cell analysis—a survey of deep learning for scRNA-seq analysis //Briefings in Bioinformatics. – 2022. – Т. 23. – №. 1. – Article bbab531.
154. Tian T. et al. Clustering single-cell RNA-seq data with a model-based deep learning approach //Nature Machine Intelligence. – 2019. – Т. 1. – №. 4. – С. 191-198.
155. Tian T. et al. Clustering single-cell RNA-seq data with a model-based deep learning approach //Nature Machine Intelligence. – 2019. – Т. 1. – №. 4. – С. 191-198.
156. Bellman R. Dynamic Programming. – Princeton: Princeton University Press, 2013. – 392 p.
157. Kleinberg J. An impossibility theorem for clustering //Advances in Neural Information Processing Systems. – 2002. – Т. 15. – С. 463-470.
158. Kiselev V. Y., Andrews T. S., Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data //Nature Reviews Genetics. – 2019. – Т. 20. – №. 5. – С. 273-282.
159. Yu D., Huber W., Vitek O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size //Bioinformatics. – 2013. – Т. 29. – №. 10. – С. 1275-1282.
160. Robinson M. D., Smyth G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data //Biostatistics. – 2008. – Т. 9. – №. 2. – С. 321-332.
161. Hafemeister C., Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression //Genome Biology. – 2019. – Т. 20. – Article 296.
162. Stuart T. et al. Comprehensive Integration of Single-Cell Data //Cell. – 2019. – Т. 177. – №. 7. – С. 1888-1902.e21.
163. Kujawa T., Marczyk M., Polańska J. Influence of single-cell RNA sequencing data integration on the performance of differential gene expression analysis //Frontiers in Genetics. – 2022. – Т. 13. – Article 959987.

164. Stuart T. et al. Comprehensive integration of single-cell data //Cell. – 2019. – T. 177. – №. 7. – C. 1888-1902.e21.
165. Hafemeister C., Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression //Genome Biology. – 2019. – T. 20. – Article 296.
166. Aran D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage //Nature Immunology. – 2019. – T. 20. – №. 2. – C. 163-172.
167. Martens J. H. A., Stunnenberg H. G. BLUEPRINT: mapping human blood cell epigenomes //Haematologica. – 2013. – T. 98. – №. 10. – C. 1487-1489.
168. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome //Nature. – 2012. – T. 489. – №. 7414. – C. 57-74.
169. Aran D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage //Nature Immunology. – 2019. – T. 20. – №. 2. – C. 163-172.
170. Mabbott N. A. et al. An expression atlas of human primary cells: inference of gene function from coexpression networks //BMC Genomics. – 2013. – T. 14. – Article 632.
171. Franzén O., Gan L. M., Björkegren J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data //Database. – 2019. – T. 2019. – Article baz046.
172. Zhang X. et al. CellMarker: a manually curated resource of cell markers in human and mouse //Nucleic Acids Research. – 2019. – T. 47. – №. D1. – C. D721-D728.
173. Franzén O., Björkegren J. L. M. alona: a web server for single-cell RNA-seq analysis //Bioinformatics. – 2020. – T. 36. – №. 12. – C. 3910-3912.
174. Mabbott N. A. et al. An expression atlas of human primary cells: inference of gene function from coexpression networks //BMC Genomics. – 2013. – T. 14. – Article 632.

175. Martens J. H. A., Stunnenberg H. G. BLUEPRINT: mapping human blood cell epigenomes //Haematologica. – 2013. – Т. 98. – №. 10. – С. 1487-1489.
176. Franzén O., Gan L. M., Björkegren J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data //Database. – 2019. – Т. 2019. – Article baz046.
177. Zhang X. et al. CellMarker: a manually curated resource of cell markers in human and mouse //Nucleic Acids Research. – 2019. – Т. 47. – №. D1. – С. D721-D728.
178. Szklarczyk D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets //Nucleic Acids Research. – 2019. – Т. 47. – №. D1. – С. D607-D613.
179. Hao Y. et al. Integrated analysis of multimodal single-cell data //Cell. – 2021. – Т. 184. – №. 13. – С. 3573-3587.e29.
180. Wang L. et al. scCapsNet-mask: an updated version of scCapsNet with extended applicability in functional analysis related to scRNA-seq data //BMC Bioinformatics. – 2022. – Т. 23. – Article 539.
181. Wang L. et al. scCapsNet-mask: an updated version of scCapsNet with extended applicability in functional analysis related to scRNA-seq data //BMC Bioinformatics. – 2022. – Т. 23. – Article 539.
182. Wang L. et al. scCapsNet-mask: an updated version of scCapsNet with extended applicability in functional analysis related to scRNA-seq data //BMC Bioinformatics. – 2022. – Т. 23. – Article 539.
183. Weiler P. et al. A guide to trajectory inference and RNA velocity //bioRxiv. – 2021. – doi: 10.1101/2021.12.22.473434.
184. Weiler P. et al. A guide to trajectory inference and RNA velocity //Single Cell Transcriptomics: Methods and Protocols. – New York: Springer, 2022. – С. 269-292.

185. Bergen V. et al. Generalizing RNA velocity to transient cell states through dynamical modeling //Nature Biotechnology. – 2020. – T. 38. – №. 12. – C. 1408-1414.
186. Weiler P. et al. A guide to trajectory inference and RNA velocity //Single Cell Transcriptomics: Methods and Protocols. – New York: Springer, 2022. – C. 269-292.
187. Wolf F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells //Genome Biology. – 2019. – T. 20. – Article 59.
188. Saelens W. et al. A comparison of single-cell trajectory inference methods //Nature Biotechnology. – 2019. – T. 37. – №. 5. – C. 547-554.
189. Street K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics //BMC Genomics. – 2018. – T. 19. – Article 477.
190. Trapnell C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells //Nature Biotechnology. – 2014. – T. 32. – №. 4. – C. 381-386.
191. Qiu X. et al. Single-cell mRNA quantification and differential analysis with Census //Nature Methods. – 2017. – T. 14. – №. 3. – C. 309-315.
192. Qiu X. et al. Reversed graph embedding resolves complex single-cell trajectories //Nature Methods. – 2017. – T. 14. – №. 10. – C. 979-982.
193. Pereira W. J. et al. Asc-Seurat: analytical single-cell Seurat-based web application //BMC Bioinformatics. – 2021. – T. 22. – Article 556.
194. Chang W. et al. shiny: Web Application Framework for R. R package version 1.7.4.9002. – 2023. – URL: https://shiny.posit.co/.
195. Hao Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis //Nature Biotechnology. – 2023. – doi: 10.1038/s41587-023-01767-y.
196. Saelens W. et al. A comparison of single-cell trajectory inference methods //Nature Biotechnology. – 2019. – T. 37. – №. 5. – C. 547-554.

197. Durinck S. et al. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt //Nature Protocols. – 2009. – Т. 4. – №. 8. – С. 1184-1191.
198. Trapnell C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells //Nature Biotechnology. – 2014. – Т. 32. – №. 4. – С. 381-386.
199. Qiu X. et al. Single-cell mRNA quantification and differential analysis with Census //Nature Methods. – 2017. – Т. 14. – №. 3. – С. 309-315.
200. Qiu X. et al. Reversed graph embedding resolves complex single-cell trajectories //Nature Methods. – 2017. – Т. 14. – №. 10. – С. 979-982.
201. Street K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics //BMC Genomics. – 2018. – Т. 19. – Article 477.
202. Perraudeau F. et al. Bioconductor workflow for single-cell RNA sequencing: Normalization, dimensionality reduction, clustering, and lineage inference //F1000Research. – 2017. – Т. 6. – Article 1158.
203. Cole M. B. et al. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq //Cell Systems. – 2019. – Т. 8. – №. 4. – С. 315-328.e8.
204. Risso D. et al. A general and flexible method for signal extraction from single-cell RNA-seq data //Nature Communications. – 2018. – Т. 9. – Article 284.
205. Risso D. et al. ClusterExperiment and RSEC: A Bioconductor package and framework for clustering of single-cell and other large gene expression datasets //PLoS Computational Biology. – 2018. – Т. 14. – №. 9. – Article e1006378.
206. Saelens W. et al. A comparison of single-cell trajectory inference methods //Nature Biotechnology. – 2019. – Т. 37. – №. 5. – С. 547-554.
207. Saelens W. et al. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools //bioRxiv. – 2018. – doi: 10.1101/276907.
208. La Manno G. et al. RNA velocity of single cells //Nature. – 2018. – Т. 560. – №. 7719. – С. 494-498.

209. Bergen V., Soldatov R. A., Kharchenko P. V., Theis F. J. RNA velocity—current challenges and future perspectives //Molecular Systems Biology. – 2021. – T. 17. – №. 8. – Article e10282.
210. Bergen V. et al. Generalizing RNA velocity to transient cell states through dynamical modeling //Nature Biotechnology. – 2020. – T. 38. – №. 12. – C. 1408-1414.
211. Bergen V., Soldatov R. A., Kharchenko P. V., Theis F. J. RNA velocity—current challenges and future perspectives //Molecular Systems Biology. – 2021. – T. 17. – №. 8. – Article e10282.
212. Bergen V., Soldatov R. A., Kharchenko P. V., Theis F. J. RNA velocity—current challenges and future perspectives //Molecular Systems Biology. – 2021. – T. 17. – №. 8. – Article e10282.
213. Lederer A. R., La Manno G. The emergence and promise of single-cell temporal-omics approaches //Current Opinion in Biotechnology. – 2020. – T. 63. – C. 70-78.
214. La Manno G. et al. RNA velocity of single cells //Nature. – 2018. – T. 560. – №. 7719. – C. 494-498.
215. Bergen V. et al. Generalizing RNA velocity to transient cell states through dynamical modeling //Nature Biotechnology. – 2020. – T. 38. – №. 12. – C. 1408-1414.
216. Aibar S. et al. SCENIC: single-cell regulatory network inference and clustering //Nature Methods. – 2017. – T. 14. – №. 11. – C. 1083-1086.
217. Schwalie P. C. et al. A stromal cell population that inhibits adipogenesis in mammalian fat depots //Nature. – 2018. – T. 559. – №. 7712. – C. 103-108.
218. Burl R. B. et al. Deconstructing adipogenesis induced by β3-adrenergic receptor activation with single-cell expression profiling //Cell Metabolism. – 2018. – T. 28. – №. 2. – C. 300-309.e4.
219. Faure L. et al. Single cell RNA sequencing identifies early diversity of sensory neurons forming via bi-potential intermediates //Nature Communications. – 2020. – T. 11. – Article 4175.

220. Acosta J. R. et al. Single cell transcriptomics suggest that human adipocyte progenitor cells constitute a homogeneous cell population //Stem Cell Research & Therapy. – 2017. – T. 8. – Article 250.
221. Liu X. et al. Single-cell RNA-seq of cultured human adipose-derived mesenchymal stem cells //Scientific Data. – 2019. – T. 6. – Article 190031.
222. Kameneva P. et al. Single-cell transcriptomics of human embryos identifies multiple sympathoblast lineages with potential implications for neuroblastoma origin //Nature Genetics. – 2021. – T. 53. – №. 5. – C. 694-706.
223. Muto Y. et al. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney //Nature Communications. – 2021. – T. 12. – Article 2190.
224. Fang R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC //Nature Communications. – 2021. – T. 12. – Article 1337.
225. Franzén O., Gan L. M., Björkegren J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data //Database. – 2019. – T. 2019. – Article baz046.
226. Zhang X. et al. CellMarker: a manually curated resource of cell markers in human and mouse //Nucleic Acids Research. – 2019. – T. 47. – №. D1. – C. D721-D728.
227. Szklarczyk D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets //Nucleic Acids Research. – 2019. – T. 47. – №. D1. – C. D607-D613.
228. Aibar S. et al. SCENIC: single-cell regulatory network inference and clustering //Nature Methods. – 2017. – T. 14. – №. 11. – C. 1083-1086.
229. Kalinina N. et al. Characterization of secretomes provides evidence for adipose-derived mesenchymal stromal cells subtypes //Stem Cell Research & Therapy. – 2015. – T. 6. – Article 221.

230. De Wever O., Demetter P., Mareel M., Bracke M. Stromal myofibroblasts are drivers of invasive cancer growth //International Journal of Cancer. – 2008. – T. 123. – №. 10. – C. 2229-2238.
231. Pho M. et al. Cofilin is a marker of myofibroblast differentiation in cells from porcine aortic cardiac valves //American Journal of Physiology-Heart and Circulatory Physiology. – 2008. – T. 294. – №. 4. – C. H1767-H1778.
232. Bühling F. et al. Pivotal role of cathepsin K in lung fibrosis //The American Journal of Pathology. – 2004. – T. 164. – №. 6. – C. 2203-2216.
233. Falconer A. M. D. et al. Collagenolytic matrix metalloproteinases antagonize proteinase-activated receptor-2 activation, providing insights into extracellular matrix turnover //Journal of Biological Chemistry. – 2019. – T. 294. – №. 26. – C. 10266-10277.
234. Bauer A., Habior A. Concentration of Serum Matrix Metalloproteinase-3 in Patients With Primary Biliary Cholangitis //Frontiers in Immunology. – 2022. – T. 13. – Article 885229.
235. Liu G. et al. Fibulin-1c regulates transforming growth factor–β activation in pulmonary tissue fibrosis //JCI Insight. – 2019. – T. 4. – №. 16. – Article e124529.
236. Ott L. E. et al. Fibroblast Migration Is Regulated by Myristoylated Alanine-Rich C-Kinase Substrate (MARCKS) Protein //PLoS One. – 2013. – T. 8. – №. 6. – Article e66512.
237. Zhou Y. et al. Chitinase 3-like 1 suppresses injury and promotes fibroproliferative responses in Mammalian lung fibrosis //Science Translational Medicine. – 2014. – T. 6. – №. 240. – Article 240ra76.
238. Wang J. et al. Macrophage-derived GPNMB trapped by fibrotic extracellular matrix promotes pulmonary fibrosis //Communications Biology. – 2023. – T. 6. – Article 136.
239. Wang P. W. et al. Characterization of the Roles of Vimentin in Regulating the Proliferation and Migration of HSCs during Hepatic Fibrogenesis //Cells. – 2019. – T. 8. – №. 10. – Article 1184.

240. Paolini C. et al. PDGF/PDGFR: A Possible Molecular Target in Scleroderma Fibrosis //International Journal of Molecular Sciences. – 2022. – Т. 23. – №. 7. – Article 3904.
241. Vervoort S. J. et al. SOX4 mediates TGF-β-induced expression of mesenchymal markers during mammary cell epithelial to mesenchymal transition //PLoS One. – 2013. – Т. 8. – №. 1. – Article e53238.
242. Russo J. M. et al. Distinct temporal-spatial roles for rho kinase and myosin light chain kinase in epithelial purse-string wound closure //Gastroenterology. – 2005. – Т. 128. – №. 4. – С. 987-1001.
243. Borkham-Kamphorst E. et al. The anti-fibrotic effects of CCN1/CYR61 in primary portal myofibroblasts are mediated through induction of reactive oxygen species resulting in cellular senescence, apoptosis and attenuated TGF-β signaling //Biochimica et Biophysica Acta (BBA)-Molecular Cell Research. – 2014. – Т. 1843. – №. 5. – С. 902-914.
244. Iwasaki T. et al. Diphosphorylated MRLC is required for organization of stress fibers in interphase cells and the contractile ring in dividing cells //Cell Structure and Function. – 2001. – Т. 26. – №. 6. – С. 677-683.
245. Kai F. B., Fawcett J. P., Duncan R. Synaptopodin-2 induces assembly of peripheral actin bundles and immature focal adhesions to promote lamellipodia formation and prostate cancer cell migration //Oncotarget. – 2015. – Т. 6. – №. 13. – С. 11162-11176.
246. Boczkowska M. et al. How Leiomodin and Tropomodulin use a common fold for different actin assembly functions //Nature Communications. – 2015. – Т. 6. – Article 8314.
247. Szklarczyk D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets //Nucleic Acids Research. – 2019. – Т. 47. – №. D1. – С. D607-D613.
248. Yates A. D. et al. Ensembl 2020 //Nucleic Acids Research. – 2020. – Т. 48. – №. D1. – С. D682-D688.

249. Li Z. et al. LncExpDB: an expression database of human long non-coding RNAs //Nucleic Acids Research. – 2021. – Т. 49. – №. D1. – С. D962-D968.
250. Hou M. et al. LocExpress: a web server for efficiently estimating expression of novel transcripts //BMC Genomics. – 2016. – Т. 17. – №. Suppl 13. – Article 847.
251. Buniello A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019 //Nucleic Acids Research. – 2019. – Т. 47. – №. D1. – С. D1005-D1012.
252. Du C. et al. The long non-coding RNA LINC01705 regulates the development of breast cancer by sponging miR-186-5p to mediate TPR expression as a competitive endogenous RNA //Frontiers in Genetics. – 2020. – Т. 11. – Article 779.
253. Zhu L., Wang X. Integrative network analysis identified master regulatory long non-coding RNAs underlying the squamous subtype of pancreatic ductal adenocarcinoma //2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). – IEEE, 2020. – С. 2936-2942.
254. Yang S. et al. Long noncoding RNA ERLR mediates epithelial-mesenchymal transition of retinal pigment epithelial cells and promotes experimental proliferative vitreoretinopathy //Cell Death & Differentiation. – 2021. – Т. 28. – №. 8. – С. 2351-2366.
255. Rey F. et al. Role of long non-coding RNAs in adipogenesis: State of the art and implications in obesity and obesity-associated diseases //Obesity Reviews. – 2021. – Т. 22. – №. 7. – Article e13203.